JCTC Journal of Chemical Theory and Computation

# Helix-Hairpin Transitions of a Designed Peptide Studied by a Generalized-Ensemble Simulation

Satoru G. Itoh,*[,†] Atsuo Tamura,[‡] and Yuko Okamoto*[,†]

*Department of Physics, School of Science,*
*Nagoya University, Nagoya, Aichi 464-8602, Japan, and*
*Department of Chemistry, Graduate School of Science,*
*Kobe University, Kobe 657-8501, Japan*

Received November 10, 2009

**Abstract:** It was recently reported that the designed peptide, whose sequence is INYWLAHAKAGYIVHWTA, has both of the two structures, α-helix and *β*-hairpin, in aqueous solution. However, the detailed transformation between these two structures is still unclear. In order to study this transformation, we applied a generalized-ensemble simulation to the designed peptide in aqueous solution and deduced the pathways of the designed peptide between α-helix structures and *β*-hairpin structures.

## 1. Introduction

Many proteins transform their tertiary structures to carry out their proper biological functions in vivo. Transformations of secondary structures play a role in such tertiary-structure-transformation processes. Therefore, it is important to understand the detailed transformations of the secondary structures. However, it is difficult to investigate the detailed transformations in experiments. Computer simulations are now widely used for such studies.

In order to investigate protein structures and their structure changes, we must realize effective samplings in the conformational space. In the conventional canonical-ensemble simulations,[1−6] however, it is difficult to achieve this in complex systems such as proteins. This is because the usual canonical-ensemble simulations tend to get trapped in a few of the many local-minimum-energy states. To overcome these difficulties, the generalized-ensemble algorithms have been proposed (for reviews, see, for instance, refs 7 and 8).

We recently developed a new generalized-ensemble algorithm, which is referred to as the multicanonical−multioverlap algorithm.[9,10] This algorithm combines the advantages of the multicanonical[11−14] and multioverlap algorithms.[15−17] The multicanonical algorithm is one of the most well-known generalized-ensemble algorithms and realizes effective samplings in the conformational space. The multioverlap algorithm also samples effectively the vicinity of specific conformations. The multicanonical−multioverlap algorithm realizes effective samplings in the conformational space more than these two algorithms.[9,10] It is useful to apply this algorithm to protein systems in order to investigate detailed protein-structure changes.[18]

One of the present authors recently designed a new peptide, whose sequence is INYWLAHAKAGYIVHWTA, in order to understand stabilizing mechanisms of protein structures.[19] This peptide has both an α-helix structure (PDB ID code 2DX3) and a *β*-sheet structure (PDB ID code 2DX4) in aqueous solution. However, the detailed transformation between these two structures is still unclear. Therefore, we applied the multicanonical−multioverlap algorithm to this designed peptide to study the transformation between α-helix and *β*-sheet structures.

In section 2, we summarize the multicanonical−multioverlap algorithm. In section 3, we describe the details of the multicanonical−multioverlap molecular dynamics (MD) simulations that we performed. We present their results in section 4. Section 5 is devoted to conclusions.

## 2. Methods

**2.1. Multicanonical−Multioverlap Algorithm.** In the multicanonical−multioverlap algorithm,[9,10] by employing the non-Boltzmann weight factor $W_{\mathrm{mco}}$, which we refer to as the multicanonical−multioverlap weight factor, a uniform probability distribution with respect to the potential energy and a dihedral-angle distance is obtained:

$$P_{\mathrm{mco}}(E,d) = n(E,d)W_{\mathrm{mco}}(E,d) \equiv \text{constant} \qquad (1)$$

where $E$ is the potential energy and $d$ is a dihedral-angle distance. The dihedral-angle distance is defined by

$$d = \frac{1}{n\pi}\sum_{i=1}^{n} d(v_i, v_i^0) \qquad (2)$$

where $n$ is the total number of dihedral angles, $v_i$ is the dihedral angle $i$, and $v_i^0$ is the dihedral angle $i$ of the reference conformation. The distance $d(v_i,v_i^0)$ between two dihedral angles is given by

$$d(v_i, v_i^0) = \min(|v_i - v_i^0|, 2\pi - |v_i - v_i^0|) \qquad (3)$$

* Corresponding author e-mail:itoh@tb.phys.nagoya-u.ac.jp (S.G.I.), okamoto@phys.nagoya-u.ac.jp (Y.O.).

[†] Nagoya University.

[‡] Kobe University.

10.1021/ct9005932
Published on Web 03/15/2010

The multicanonical-multioverlap weight factor at a constant temperature $T_0$ can be written as

$$W_{\rm mco}(E, d) = {\rm e}^{-\beta_0 E_{\rm mco}(E,d)} \tag{4}$$

where $\beta_0$ is defined by $\beta_0 = 1/k_B T_0$ ($k_B$ is the Boltzmann constant) and $E_{\rm mco}(E,d)$ is the multicanonical−multioverlap potential energy. Equation 1 implies that multicanonical−multioverlap simulations realize a random walk in the two-dimensional potential-energy and dihedral-angle-distance space and are able to effectively sample the conformational space. Accordingly, we can obtain accurate free-energy landscapes of protein systems and estimate folding pathways and the transition states among the specific conformations.[9,10] We remark that we employed the MD version of the multicanonical−multioverlap algorithm[18] in this article.

**2.2. Reweighting Techniques.** The results from a multicanonical−multioverlap simulation can be analyzed by the reweighting techniques. Suppose that we have determined the multicanonical−multioverlap potential energy $E_{\rm mco}$ in eq 4 at a constant temperature $T_0$ and that we performed the simulation at this temperature. The expectation value of a physical quantity $A$ at any temperature $T$ is calculated from

$$\langle A \rangle_T = \frac{\sum\limits_{E,d} A(E, d)\, n(E, d)\, {\rm e}^{-\beta E}}{\sum\limits_{E,d} n(E, d)\, {\rm e}^{-\beta E}} \tag{5}$$

where the best estimate of the density of states is given by the single-histogram reweighting techniques[20,21] (see eq 1):

$$n(E, d) = \frac{N_{\rm mco}(E, d)}{W_{\rm mco}(E, d)} \tag{6}$$

and $N_{\rm mco}(E,d)$ is the histogram of the probability distribution that was obtained by the multicanonical−multioverlap simulation.

We can also calculate the free energy (or the potential of mean force) with appropriate reaction coordinates. For example, the free energy $F(\xi_1,\xi_2;T)$ with reaction coordinates $\xi_1$ and $\xi_2$ at temperature $T$ is given by

$$F(\xi_1, \xi_2;T) = -k_B T \ln P_c(\xi_1, \xi_2;T) \tag{7}$$

where $P_c(\xi_1,\xi_2;T)$ is the reweighted canonical probability distribution of $\xi_1$ and $\xi_2$ and given by (see eq 5)

$$P_c(\xi_1, \xi_2;T) = \frac{\sum\limits_{E,d} N_{\rm mco}(\xi_1, \xi_2;E, d)\, {\rm e}^{\beta_0 E_{\rm mco}(E,d) - \beta E}}{\sum\limits_{\xi_1,\xi_2,E,d} N_{\rm mco}(\xi_1, \xi_2;E, d)\, {\rm e}^{\beta_0 E_{\rm mco}(E,d) - \beta E}} \tag{8}$$

and $N_{\rm mco}(\xi_1,\xi_2;E,d)$ is the histogram of the probability distribution that was obtained from the multicanonical−multioverlap simulation.

## 3. Computational Details

It was reported that the designed peptide, whose sequence is INYWLAHAKAGYIVHWTA, has the two characteristic structures, α-helix and $\beta$-hairpin, coexisting in a pH 4.5 solution in experiments.[19] In our simulation, the histidine residues of the designed peptide were protonated in order to conform our simulation conditions to the low pH ones in the experiment. The force field that we adopted is the CHARMM 22 parameter set.[22] We employed the Generalized Born/Surface Area (GB/SA) model[23−25] as an implicit solvent model.

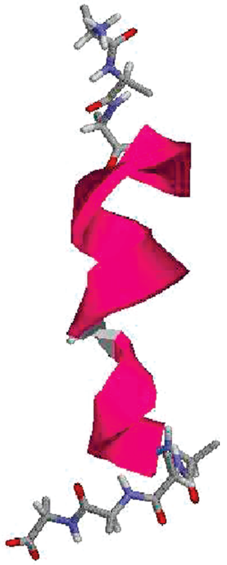


***Figure 1.*** Reference conformation in our multicanonical−multioverlap simulation. The N terminus and the C terminus are on the top side and on the bottom side, respectively. The figure was created with RasMol.[28]

***Table 1.*** Backbone Dihedral Angles of the Reference Conformation in Figure 1

| | angle (deg) | | angle (deg) |
|---|---|---|---|
| $\phi_2$ | −153.7 | $\phi_{10}$ | −76.4 |
| $\psi_2$ | 153.0 | $\psi_{10}$ | −30.0 |
| $\phi_3$ | −96.8 | $\phi_{11}$ | −95.4 |
| $\psi_3$ | −51.5 | $\psi_{11}$ | 15.5 |
| $\phi_4$ | −67.6 | $\phi_{12}$ | −70.7 |
| $\psi_4$ | −44.6 | $\psi_{12}$ | −3.9 |
| $\phi_5$ | −83.9 | $\phi_{13}$ | −69.8 |
| $\psi_5$ | 22.6 | $\psi_{13}$ | −8.9 |
| $\phi_6$ | −125.9 | $\phi_{14}$ | −64.5 |
| $\psi_6$ | −56.3 | $\psi_{14}$ | −46.4 |
| $\phi_7$ | −83.3 | $\phi_{15}$ | −84.4 |
| $\psi_7$ | −29.0 | $\psi_{15}$ | −45.6 |
| $\phi_8$ | −85.3 | $\phi_{16}$ | −64.1 |
| $\psi_8$ | −1.4 | $\psi_{16}$ | −83.8 |
| $\phi_9$ | −101.3 | $\phi_{17}$ | −161.2 |
| $\psi_9$ | −50.2 | $\psi_{17}$ | 143.1 |

In multicanonical−multioverlap simulations, we must have a reference conformation. We adopted the conformation in Figure 1 as the reference conformation in this article. This conformation was obtained by minimizing the α-helix structure of the designed peptide which corresponds to the model 1 structure in the 2DX3 PDB file.[19] We list the backbone-dihedral angles of the reference conformation in Table 1, and these values were employed as the reference dihedral angles $v_i^0$ of the dihedral-angle distance in eq 2. Hence, we took into account only the backbone-dihedral angles $\phi$ (the rotation angles around the N−C$_\alpha$ bonds) and $\psi$ (the rotation angles around the C$_\alpha$−C bonds) of the residues 2−17 of the designed peptide as the reference dihedral angles in our simulations. We remark that, although we employed the α-helix structure as a reference

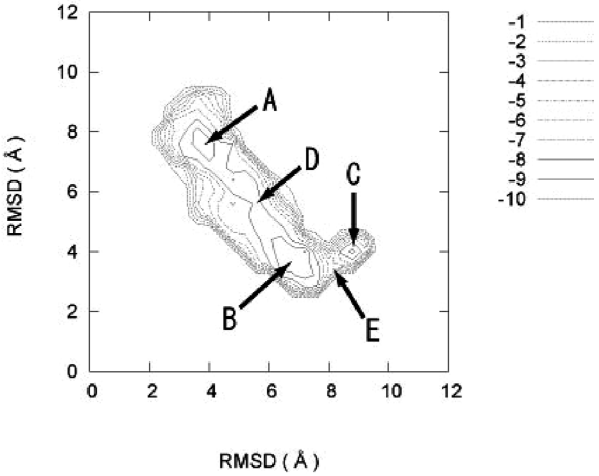


***Figure 2.*** Free-energy landscape obtained from the results of the multicanonical−multioverlap MD simulation at 300 K. The abscissa and the ordinate are the RMSDs of the backbone $C_\alpha$ atoms with respect to the $\alpha$-helix structure in Figure 1 and the hairpin structure in Figure 3, respectively. Contour lines are drawn every 1 kcal/mol. The labels A, B, and C locate the free-energy local-minimum states. The labels D and E stand for the saddle points between A and B and between B and C, respectively.

conformation, results of multicanonical−multioverlap simulations are independent of selection of reference conformations as long as sampling is performed sufficiently.[9,10]

The multicanonical−multioverlap weight factor was first determined so that a free random walk was realized in the two-dimensional energy-overlap space. As for the potential-energy random walk, this weight factor covered the temperature range from 300 to 600 K (see Supporting Information for the determination of the multicanonical−multioverlap weight factor). We then performed a multicanonical−multioverlap production run, in which the time step was 0.5 fs, for 44.5 ns after an equilibration of 0.5 ns. The initial conformation of the designed peptide for the production run was a random-coil conformation.

## 4. Results and Discussion

We show the free-energy landscape at 300 K in Figure 2. The free-energy landscape was obtained from the results of the multicanonical−multioverlap MD simulation by the reweighting techniques in eqs 7 and 8. The abscissa is the root-mean-square distance (RMSD) of the backbone $C_\alpha$ atoms with respect to the reference conformation in Figure 1. The RMSD with respect to the reference conformation is defined by

$$\mathrm{RMSD} = \min\left[\sqrt{\frac{1}{N}\sum_i (\mathbf{q}_i - \mathbf{q}_i^0)^2}\right] \tag{9}$$

where $N$ is the number of atoms, $\{\mathbf{q}_i{}^0\}$ are the coordinates of the reference conformation, and the minimization is over the rigid translations and rigid rotations of the coordinates of the conformation $\{\mathbf{q}_i\}$ with respect to the center of geometry. The ordinate is the RMSD with respect to the hairpin structure of the designed peptide in Figure 3. This structure was obtained by minimizing the hairpin structure of the model 1 structure in the 2DX4 PDB file.[19] Three free-energy local-minimum states and two transition states among these local-minimum states are identified in Figure 2. We label the three local-minimum states as A, B, and C; the transition state between the local-minimum states A and B as D; and the transition state between the local-minimum states B and C as E. These local-minimum states were not completely coincident with the experimental conformations in Figures 1 and 3. In other words, RMSDs between the local-minimum states and the experimental conformations were more than 2 Å at 300 K. This is because conformations whose RMSDs were less than 2 Å had high potential energy with the CHARMM force field. We remark that the free-energy landscape with respect to the dihedral-angle distance $d$ is described in the Supporting Information as a reference.

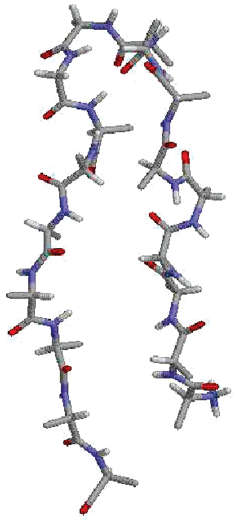


***Figure 3.*** The hairpin structure of the designed peptide. The N-terminus and the C-terminus are on the right-hand side and on the left-hand side, respectively. The figure was created with RasMol.[28]

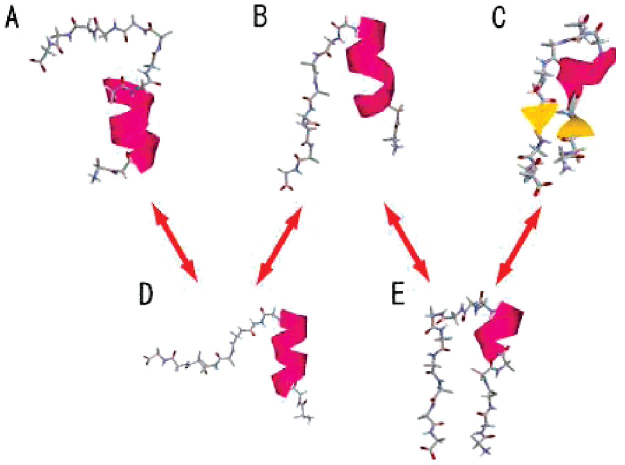


***Figure 4.*** Typical structures in the corresponding local-minimum states in Figure 2. The N-terminus and the C-terminus are on the right-hand side and on the left-hand side, respectively. The arrows indicate possible pathways of the designed peptide from $\alpha$-helix structure to hairpin-like structure. The figures were created with RasMol.[28]

In Figure 4, we show typical conformations of the designed peptide in each local-minimum state (see the Supporting Information for how these typical conformations were obtained). In the free-energy local-minimum state A, the designed peptide

***Table 2.*** Expectation Values of the α-Helix Content and the End-to-End Distance and Values of Free Energy in the Local-Minimum States in Figure 2[a]

| | A | B | C | Figure 1 | Figure 3 |
|---|---|---|---|---|---|
| α-helicity (%) | 27.1 (2.2) | 29.5 (5.0) | 22.0 (0.3) | 38.9 | 0.0 |
| end-to-end distance (Å) | 23.8 (0.8) | 13.2 (0.9) | 4.5 (0.3) | 25.7 | 7.0 |
| free energy (kcal/mol) | −9.2 (0.4) | −8.9 (0.3) | −8.5 (1.5) | | |

[a] The corresponding values for the conformations in Figures 1 and 3 are also given as reference. Errors were estimated by the jackknife method.[29−31]

has α-helix structures with an extended C-terminus. In B, this extended C-terminus is close to the N-terminus. The extended C-terminus forms a $\beta$-ladder with the N-terminus in the local-minimum state C. These hairpin-like structures are similar to the structure in Figure 3. We also show typical conformations in transition states in Figure 4. In the transition state D between the free-energy local-minimum states A and B, the extended C-terminus is almost vertical with respect to the helix axis. Because the C-terminal parts are extended to the upper side of the α-helix structure in A and to the lower side of the α-helix structure in B such as the conformations in Figure 4, the conformation in the transition state D is very reasonable as an intermediate conformation between those in A and B. The structures in the transition state E between the local-minimum states B and C have an extended N-terminus, and this extended N-terminus is created by breaking the corresponding part of the α-helix structure in B. We remark that, although we determined the conformations in the transition states in Figure 4 simply as shown in the Supporting Information, there is a useful analysis method to obtain more accurate transition states among free-energy local minima that was presented in ref 26.

We list in Table 2 the expectation values of the α-helicity and end-to-end distance and the values of the free energy in the local-minimum states in Figure 2. Moreover, we listed the α-helicity and end-to-end distance of the conformations in Figures 1 and 3 as references. Here, an α-helicity of a conformation is defined by

$$\frac{N_\alpha}{N_{\text{res}}} \tag{10}$$

where $N_\alpha$ is the number of residues that have helix structures and $N_{\text{res}}$ is the number of the residues of the peptide ($N_{\text{res}} = 18$ for the designed peptide). We employed the definition of secondary structure of proteins (DSSP) criteria[27] to determine the secondary structures of the peptide. The α-helicity in the local-minimum state C is small in comparison with those in the local-minimum states A and B from this table. This is because the structures in C have a short extended N-terminus by breaking apart of N-terminal α-helix structures in A and B as mentioned above. Furthermore, the average end-to-end distance gets smaller when the state is transferred from A to B or from B to C. The expectation value of the end-to-end distance in the local-minimum state C is close to the end-to-end distance obtained from the experimental conformation in Figure 3.

From Figures 2 and 4 and Table 2, we deduce the transformation process of secondary structures of the designed peptide between α-helix structures and hairpin-like structures as follows. Stage 1: The designed peptides have an α-helix structure with an extended C-terminus such as the structure in A in Figure 4. Stage 2: The extended C-terminus is bent (transition state D) and comes close to the N-terminus like the structure in B. Stage 3: A part of the α-helix on the N-terminus is broken, and the extended N-terminus is formed (transition state E). Stage 4: The extended C-terminus and the extended N-terminus create the $\beta$-ladder such as the structure in the local-minimum state C. These pathways are summarized in Figure 4 (see the arrows).

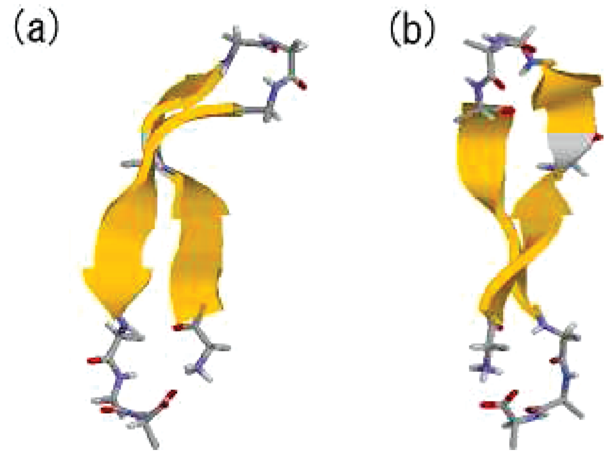


***Figure 5.*** A $\beta$-hairpin structure in the free-energy local-minimum state C in Figure 2. a and b correspond to the same conformation viewed from different angles. The figures were created with RasMol.[28]

The $\beta$-hairpin structures of the designed peptide were observed in an experiment.[19] Although the typical conformations in the local-minimum did not have complete $\beta$-hairpin structures in our simulation, complete $\beta$-hairpin structures did exist during the simulation, where the conformations were in C. Figure 5 shows the complete $\beta$-hairpin structure found in the free-energy local-minimum state C. This structure is formed by breaking the remaining α-helix structure of the typical conformation in C.

## 5. Conclusions

It is important to understand the transformations of the secondary structures of proteins. This is because the protein functions are closely associated with the tertiary structures, and the tertiary structure changes are caused by the secondary structure changes. The typical secondary structures for the proteins are α-helix and $\beta$-sheet structures, and the transformations between α-helix and $\beta$-sheet structures are often important for the protein function. For instance, amyloidogenesis is caused by the transformations from α-helix to $\beta$-sheet structures for certain proteins.

In the present work, we calculated the free-energy landscape at 300 K for the designed peptide in aqueous solution from the results of a multicanonical−multioverlap MD simulation. The designed peptide has the two structures, α-helix and $\beta$-hairpin, coexisting in an experiment. We observed these structures in our simulation and identified intermediate structures and transition states between the two structures. We also deduced the transformation pathways of the designed peptide between α-helix structures and $\beta$-hairpin structures. In other words, it was clarified that the α-helix structure of the N-terminus is broken step by step, and the β-ladder is created between the broken parts and the extended C-terminus in the transformations from α-helix to β-hairpin structures. Therefore, we believe that our results for the transformations between α-helix and β-sheet structures will play an important role in understanding the protein functions, while our results are still with a small designed peptide.

**Acknowledgment**. The computations were performed on the computers at the Research Center for Computational Science, Institute for Molecular Science. This work was supported, in part, by the Grants-in-Aid for the Next Generation Super Computing Project, Nanoscience Program and for Scientific Research on Innovative Areas, "Fluctuations and Biological Functions," from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

**Supporting Information Available:** Computational details. Free-energy landscape with respect to the dihedral-angle distance. This information is available free of charge via the Internet at http://pubs.acs.org/.

## References

(1) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of state calculations by fast computing machines. *J. Chem. Phys.* **1953**, *21*, 1087.

(2) Hoover, W. G.; Ladd, A. J. C.; Moran, B. High-strain-rate plastic-flow studied via non-equilibrium molecular-dynamics. *Phys. Rev. Lett.* **1982**, *48*, 1818.

(3) Evans, D. J. Computer experiment for non-linear thermodynamics of couette-flow. *J. Chem. Phys.* **1983**, *78*, 3297.

(4) Nosé, S. A molecular-dynamics method for simulations in the canonical ensemble. *Mol. Phys.* **1984**, *52*, 255.

(5) Nosé, S. A unified formulation of the constant temperature molecular-dynamics methods. *J. Chem. Phys.* **1984**, *81*, 511.

(6) Hoover, W. G. Canonical dynamics - equilibrium phase-space distributions. *Phys. Rev. A* **1985**, *31*, 1695.

(7) Mitsutake, A.; Sugita, Y.; Okamoto, Y. Generalized-ensemble algorithms for molecular simulations of biopolymers. *Biopolymers* **2001**, *60*, 96.

(8) Itoh, S. G.; Okumura, H.; Okamoto, Y. Generalized-ensemble algorithms for molecular dynamics simulations. *Mol. Sim.* **2007**, *33*, 47.

(9) Itoh, S. G.; Okamoto, Y. A new generalized-ensemble algorithm: multicanonical-multioverlap algorithm. *Mol. Sim.* **2007**, *33*, 83.

(10) Itoh, S. G.; Okamoto, Y. Effective sampling in the configurational space of a small peptide by the multicanonical-multioverlap algorithm. *Phys. Rev. E* **2007**, *76*, 026705.

(11) Berg, B. A.; Neuhaus, T. Multicanonical algorithms for 1st order phase-transitions. *Phys. Lett. B* **1991**, *267*, 249.

(12) Berg, B. A.; Neuhaus, T. Multicanonical ensemble - a new approach to simulate 1st-order phase-transitions. *Phys. Rev. Lett.* **1992**, *68*, 9.

(13) Hansmann, U. H. E.; Okamoto, Y.; Eisenmenger, F. Molecular dynamics, Langevin and hybrid Monte Carlo simulations in a multicanonical ensemble. *Chem. Phys. Lett.* **1996**, *259*, 321 .

(14) Nakajima, N.; Nakamura, H.; Kidera, A. Multicanonical ensemble generated by molecular dynamics simulation for enhanced conformational sampling of peptides. *J. Phys. Chem. B* **1997**, *101*, 817.

(15) Berg, B. A.; Noguchi, H.; Okamoto, Y. Multioverlap simulations for transitions between reference configurations. *Phys. Rev. E* **2003**, *68*, 036126.

(16) Itoh, S. G.; Okamoto, Y. Multi-overlap molecular dynamics methods for biomolecular systems. *Chem. Phys. Lett.* **2004**, *400*, 308.

(17) Itoh, S. G.; Okamoto, Y. Theoretical studies of transition states by the multioverlap molecular dynamics methods. *J. Chem. Phys.* **2006**, *124*, 104103.

(18) Itoh, S. G.; Okamoto, Y. Amyloid-β(29−42) dimer formations studied by a multicanonical-multioverlap molecular dynamics simulation. *J. Phys. Chem. B* **2008**, *112*, 2767.

(19) Araki, M.; Tamura, A. Transformation of an α-helix peptide into a β-hairpin induced by addition of a fragment results in creation of a coexisting state. *Proteins* **2007**, *66*, 860.

(20) Ferrenberg, A. M.; Swendsen, R. H. New Monte Carlo technique for studying phase transitions. *Phys. Rev. Lett.* **1988**, *61*, 2635.

(21) Ferrenberg, A. M.; Swendsen, R. H. New Monte Carlo technique for studying phase transitions. *Phys. Rev. Lett.* **1989**, *63*, 1658.

(22) 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.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998**, *102*, 3586.

(23) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127.

(24) Dominy, B. N.; Brooks, C. L., III. Development of a generalized born model parametrization for proteins and nucleic acids. *J. Phys. Chem. B* **1999**, *103*, 3765.

(25) Feig, M.; Brooks, C. L., III. Evaluating CASP4 predictions with physical energy functions. *Proteins* **2002**, *49*, 232.

(26) Bolhuis, P. G.; Dellago, C.; Chandler, D. Reaction coordinates of biomolecular isomerization. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 5877.

(27) Kabsch, W.; Sander, C. Dictionary of protein secondary structure - pattern-recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577.

(28) Sayle, R. A.; Milner-White, E. J. Biomolecular Graphics for All. *Trends Biochem. Sci.* **1995**, *20*, 374.

(29) Quenouille, M. H. Notes on bias in estimation. *Biometrika* **1956**, *43*, 353.

(30) Miller, R. G. Jackknife - review. *Biometrika* **1974**, *61*, 1.

(31) Berg, B. A. *Markov Chain Monte Carlo Simulations and Their Statistical Analysis*; World Scientific: Singapore, 2004.

CT9005932

JCTC Journal of Chemical Theory and Computation

# Molecular Dynamics of Methyl Viologen-Cucurbit[*n*]uril Complexes in Aqueous Solution

Musa I. El-Barghouthi,*,† Khaleel I. Assaf,† and Abdel Monem M. Rawashdeh*,‡

*Department of Chemistry, The Hashemite University, P.O. Box 150459, Zarqa 13115, Jordan, and Department of Chemistry, Yarmouk University, Irbid 21163, Jordan*

Received November 23, 2009

**Abstract:** In this work, molecular dynamics (MD) simulations have been used to study the dynamics of the inclusion complexes of methyl viologen (MV) with cucurbit[*n*]uril, CB*n* (where $n = 6$, 7, and 8) in aqueous solution. The obtained MD trajectories were analyzed and post-processed using the Molecular Mechanics-Poisson–Boltzmann Surface Area (MM-PBSA) method to shed some light on the host–guest intermolecular forces that play a significant role in the formation of the CB inclusion complexes. MV exhibits partial inclusion into CB6 cavity, while deep inclusion was observed for the larger macrocyclic hosts with the two cationic groups interacting with the carbonyl portals. The extracted snapshots reveal an increase in the macrocycle distortion of CB6 and CB7 upon inclusion of the guest molecule. MM-PBSA calculations indicate that CB7 forms the most stable complex with MV. The host–guest electrostatic interactions are the dominant contribution to the complex stability. Furthermore, van der Waals interactions add significantly to the complex binding free energy. The Potential of Mean Force (PMF) for the host–guest distance was obtained by umbrella sampling. No energy barriers were obtained for the guest movement inside the host cavity except in the case of CB6.

## Introduction

Cucurbit[*n*]urils (CB*n*) are macrocyclic molecules consisting of *n* number of glycouril repeated units forming symmetric barrel-shaped structures with two oxygen-crowned portals and a hydrophobic interior.[1–5] Due to these structural features, CBs form inclusion complexes with a variety of guest molecules. CB6 was the first member synthesized in 1905 by Behrend et al.,[6] yet the supramolecular chemistry of CB6 was not launched until the 1980s and 1990s. Other sizes of Cucurbiturils, CB5, CB7, CB8, and CB10, were prepared and made commercially available.[7–9] CBs have been explored as supramolecular catalysts, and they play an important role in the construction of polyrotaxanes[10] and supramolecular switches,[11] the removal of contaminants such as colorants from water,[12] and in the pharmaceutical field.[13]

Many experimental studies were performed to examine the inclusion complexation of different types of molecules with CBs.[14–23] One of the guest molecules that showed strong binding with CB host molecules is methyl viologen (MV), a dication molecule, and hence it received great attention. Ong et al. also studied the complexation of an MV guest molecule with CB7. The obtained $^1$H NMR spectra showed that CB7 forms stable 1:1 inclusion complexes with the guest molecule in aqueous solution. The binding constant was measured by electronic absorption spectroscopy and was found to be $1.0 \times 10^5$ M$^{-1}$. The smaller host CB6 was also found to bind with MV but with a lower binding constant (20 M$^{-1}$). On the basis of their $^1$H NMR data, they described the complex as partial inclusion of MV into a CB6 cavity.[24] The interactions between CB7 with MV showed that a 1:1 host–guest complex in an aqueous solution was evidenced by $^1$H NMR and mass spectroscopy with a high binding constant of $2.0 \times 10^5$ M$^{-1}$, in contrast to $\beta$-cyclodextrin which shows a formation constant of about zero. This strong binding was explained by the favorable ion–dipole interac-

* Corresponding author e-mail: musab@hu.edu.jo (M.I.E.-B.), rawash@yu.edu.jo (A.M.M.R.).

† The Hashemite University.

‡ Yarmouk University.

10.1021/ct900622h © 2010 American Chemical Society
Published on Web 03/19/2010

tions between the positive charges of the guest and the portal oxygen atoms of CB7 in addition to the hydrophobic effect inside the cavity.[25] Moon and Kaifer studied the host−guest interactions between CB7 and a series of dialkyl viologens by $^1$H NMR spectroscopy. The results showed two modes of inclusion, one with short-chain viologenes, which showed inclusion for the aromatic group, and one with long-chain viologens, which showed inclusion of the terminal alkyl groups into the inner cavity of the host.[26] Recent studies have shown that CB8 forms 1:2 (guest/host) and even possibly a 1:3 inclusion complex with *N*,*N*′-dialkyl-viologens when the alkyl chain consists of more than four carbon atoms, while it forms 1:1 for the shorter chains. MV was shown to form a 1:1 complex with a $2.7 \times 10^4$ M$^{-1}$ binding constant.[27] Other studies have reported a value of $1.1 \times 10^5$ M$^{-1}$ for the binding constant of CB8 with MV.[28]

In this work, the inclusion complexation of MV with CB6, CB7, and CB8 in water will be explored using MD simulation techniques. MV is selected in this study as a model guest molecule due to the availability of numerous experimental data regarding its complexation with several CB hosts, including binding constants and complex geometries. The corresponding MD trajectories will be analyzed to measure the effect of CB cavity size on the inclusion process as well as to gain detailed information on the guest dynamics inside the CB cavity. The flexibility and the distortion of the CB macrocycle in the presence and absence of the guest molecule will be addressed. The Molecular Mechanics−Poisson−Boltzmann Surface Area (MM-PBSA) method will be used to estimate the binding free energy of each CB*n* complex. The components of the binding free energies will also be estimated and used to explore the type of host−guest interactions responsible for complex formation, which may provide further insights into the inclusion phenomenon. The variation of the free energy values or Potential of Mean Force (PMF) with the host−guest distance will be employed by means of the umbrella sampling method to understand more about the energy barriers that may exist during the inclusion process.

## Computational Methods

The X-ray structures of CB6, CB7, and CB8 were used as initial geometries.[5,7] The geometry optimization and electrostatic potential of the guest molecule were computed using *ab initio* HF/6-31G* calculations using the Gaussian 03W package.[29] The atomic charges for the guest molecule reproducing these electrostatic potentials were obtained using the RESP methodology,[30] whereas AM1-BCC charges were used for CB*n*.[31] The AMBER 8 software[32] was used throughout this work using the general force field parameter sets.[33] Each system was solvated by a cubic box of TIP3P water molecules with a closeness parameter of 10 Å.[34] Cl $^-$ anions were added when needed to neutralize the system. Periodic boundary conditions were adopted, and the Particle Mesh Ewald method (PME) was used for the treatment of long-range electrostatic interactions.[35] The nonbonded cutoff was set to 10.0 Å. Energy minimization was performed for each solvated complex using the conjugate gradient algorithm, heated up to 298 K for 60 ps, and followed by 200 ps equilibration at 298 K and 1 atm. Production runs were carried out for 5 ns; the system was coupled in the NPT ensemble to a Berendsen thermostat at 298 K and a barostat at 1 atm. A 2 fs time step with a saving of the structure every 2 ps was used, and the nonbonded pair list was updated every 25 steps.

Analysis of the obtained MD trajectories was conducted using the PTRAJ module of AMBER. For hydrogen bond analysis, a hydrogen bond cut distance ≤ 3.0 Å and angle ≥ 120° were used. Visualization of the obtained trajectories was done using the VMD program.[36]

For MM-PBSA calculations, 2500 snapshots of the unbound guest molecule, CB*n*, and their complexes were taken from their independent MD trajectories. The explicit water molecules and the added ions were removed in each snapshot. Details on estimating the binding free energy $\Delta G_{\text{bind}}$ and its components are described below.

The binding free energy $\Delta G_{\text{bind}}$ was estimated as follows:

$$\Delta G_{\text{bind}} = \Delta E_{\text{gas}} + \Delta G_{\text{solv}} - T\Delta S \qquad (1)$$

where $\Delta E_{\text{gas}}$ is the interaction energy between the guest and host in the gas phase and is given by

$$\Delta E_{\text{gas}} = \Delta E_{\text{INT}} + \Delta E_{\text{elect}} + \Delta E_{\text{vdW}} \qquad (2)$$

where $\Delta E_{\text{INT}}$ is the change in the internal energy upon complexation. $\Delta E_{\text{elect}}$ and $\Delta E_{\text{vdW}}$ represent the host−guest electrostatic and van der Waals interactions, respectively.

The solvation free energy $\Delta G_{\text{solv}}$ was estimated as the sum of electrostatic solvation free energy $\Delta G_{\text{PB}}$ and apolar solvation free energy $\Delta G_{\text{NP}}$:

$$\Delta G_{\text{solv}} = \Delta G_{\text{PB}} + \Delta G_{\text{NP}} \qquad (3)$$

$\Delta G_{\text{PB}}$ is computed in a continuum solvent using the PBSA program of AMBER 8, while the $\Delta G_{\text{NP}}$ was calculated from the solvent-accessible surface area (SASA), which was estimated by the MSMS program using a probe radius of 0.14 nm,[37] which is given by

$$\Delta G_{\text{NP}} = \gamma \cdot \text{SASA} + b \qquad (4)$$

where $\gamma = 0.00542$ kcal/(mol Å$^2$) and $b = 0.92$ kcal/mol.

The change of the solute entropy upon complexation, $T\Delta S_{\text{conf}}$, was estimated from normal-mode analysis using the NMODE module of the AMBER 8 program.

Since prediction of the solute configurational entropy contribution ($T\Delta S_{\text{conf}}$), computed by the normal-mode analysis of the NMODE module in the AMBER 8, is associated with a relatively large error,[38,39] two values of the binding free energies of the complexation process were calculated in this work: $\Delta G$, which is the complexation free energy excluding the solute configurational entropy contribution ($T\Delta S_{\text{conf}}$), and $\Delta G^*$, which includes the solute configurational entropy contribution ($T\Delta S_{\text{conf}}$). Potentials of mean force, PMF, were generated for the inclusion of the guest molecule into the host cavity. The reaction coordinate used in the umbrella sampling was the distance ($r$) between the methyl carbon atom attached to the nitrogen atom (dark spot) in MV and the center of mass of one CB carbonyl portal (gray portal), as shown in Figure 1. The values of $r$ ranged from

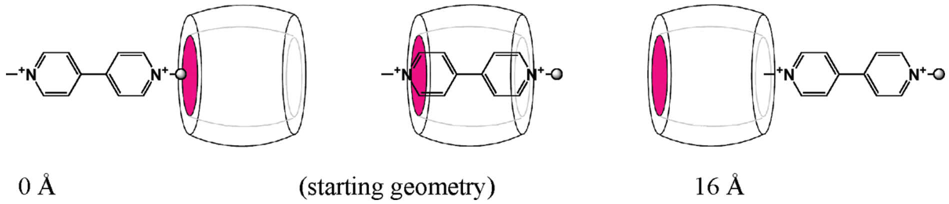


***Figure 1.*** Definition of the distance used in umbrella sampling.

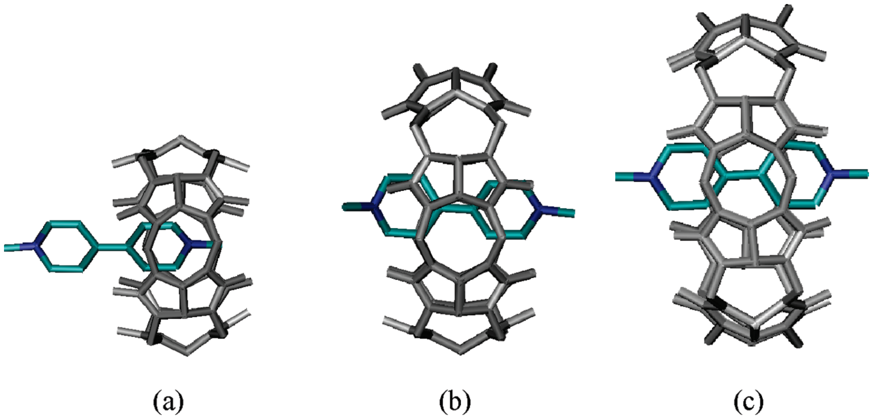


***Figure 2.*** Average structures of the studied complexes, (a) CB6/MV, (b) CB7/MV, (c) CB8/MV.

0 to 16 Å in 1.0 Å intervals. The initial distance corresponds to the inclusion of the guest molecule for CB7 and CB8, while it corresponds to the partial inclusion of the guest in the CB6 cavity. Then, the distance was varied to pull out the guest molecule from each rim of CB. Umbrella potential with a force constant of 6 kcal/mol/Å$^2$ for each position was applied. Each biasing MD simulation consists of an equilibrium run of 260 ps using a protocol similar to the conventional MD simulations described above, followed by a production run of 500 ps. The distance data for each simulation were collected in 5.0 fs intervals. The results were post-processed using the Weighted Histogram Analysis Method (WHAM).[40,41]

## Results and Discussion

Five-nanosecond MD simulations for CB6, CB7, and CB8 and their 1:1 complexes with MV were performed in water to study the dynamics of the host and guest molecules and the intermolecular forces responsible for the complex formation.

The average structures of the corresponding 1:1 complexes obtained from the 5 ns MD trajectories of the complexes are shown in Figure 2. The average structure of CB6/MV showed a partial inclusion of MV into the cavity, where one of the pyridinium rings was located inside the cavity while the other ring interacted with the surrounding water molecules. On the other hand, the average structure of CB7/MV showed a complete inclusion of the guest molecule, in harmony with earlier $^1$H NMR spectroscopic experiments, in which the $\beta$ aromatic protons of MV exhibited an upfield shift in the presence of the host. Moreover, irradiation of the methyne and methylene protons of CB7 gave rise to nuclear Overhauser effects for the α and $\beta$ aromatic protons of MV.[24] These data could only be explained by the deep inclusion of the guest molecule in the cavity of CB7.[24] This

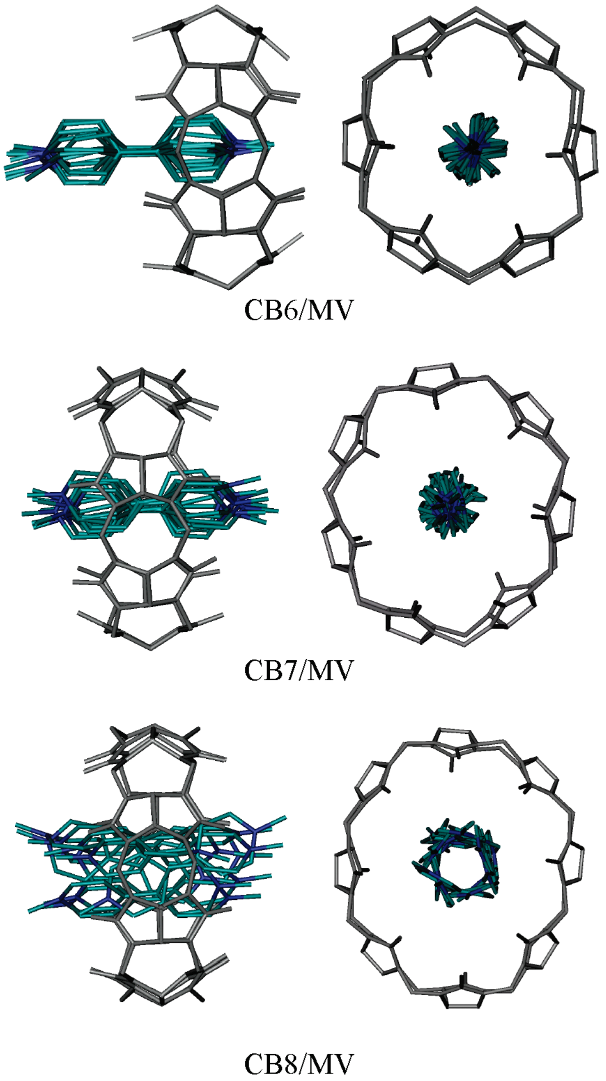
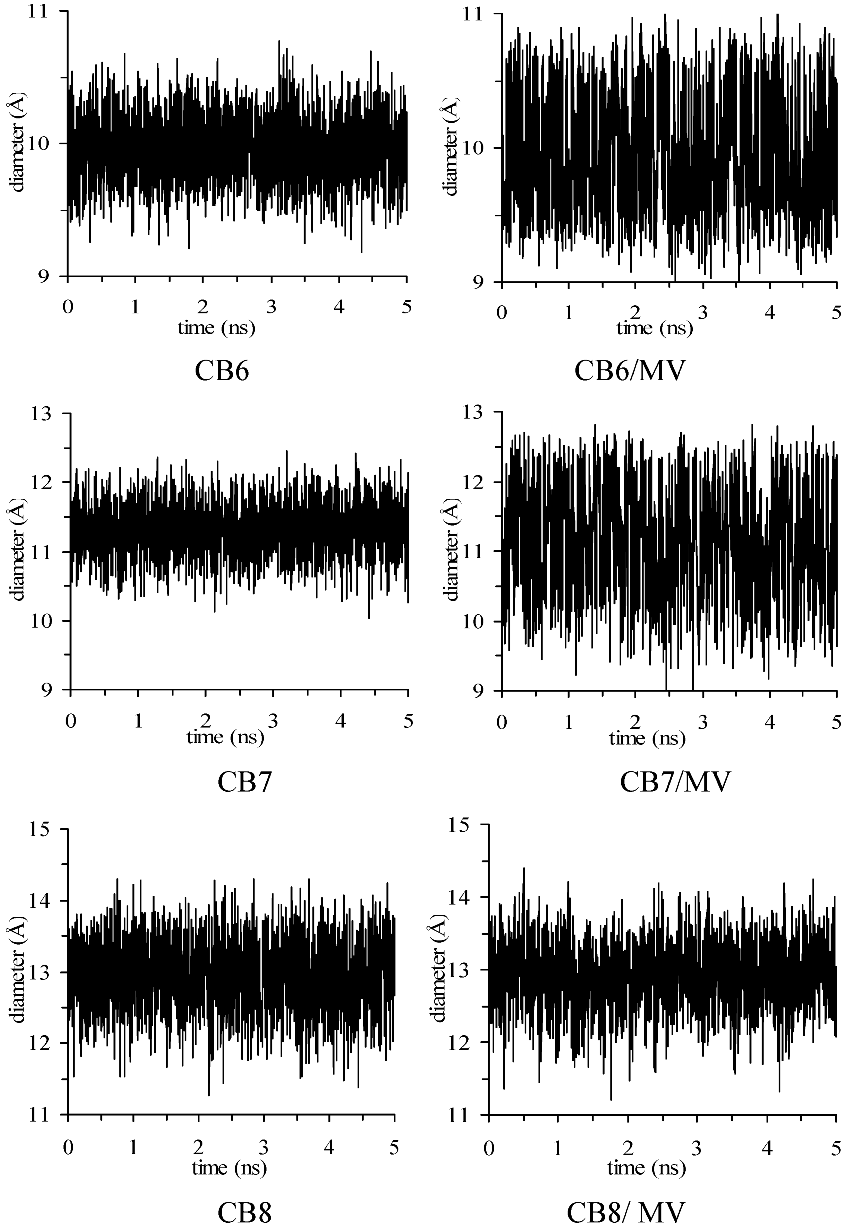


***Figure 3.*** Dynamics of 1:1 complexes, shown as a clustered molecular display for CB/MV (side and top views).



CB6 CB6/MV

CB7 CB7/MV

CB8 CB8/ MV

***Figure 4.*** CB internal diameter as a function of simulation time.

structure reveals the favorable ion−dipole interaction between the positive charge on each end of the guest molecule and the portal oxygen atoms of CB7, while the hydrophobic part of the guest molecule was accommodated in the hydrophobic cavity. Similarly, CB8/MV average geometry also depicts complete inclusion. It seems that MV has more room inside the cavity because of the larger cavity size of CB8. Although the cavity of CB8 is large enough to accommodate two aromatic ring guest molecules, Kim and co-workers found that CB8 binds a single molecule of MV.[27,28]

The guest dynamics inside the host cavity, monitored by the superposition of 10 snapshots extracted from the MD trajectory for each complex, superimposed on a representative host structure, are given in Figure 3. A first assessment of the snapshots shows restricted translational movement of the guest in and out of the cavity in all CBs. This indicates the stability of the ion−dipole interaction in CB7 and CB8 during the simulations. The electrostatic potential of the CB cavity is negative,[2] which may explain the stable location of the cationic group of the guest molecule inside the CB6 cavity. Rotational motion of MV inside the cavity was observed in all complexes. Only in CB8 was the guest molecule found to translate within the cavity due to space availability.

Figure 4 shows the internal diameter (see scheme 1) values of CB6, CB7, and CB8 as a function of time in the absence and presence of the guest molecule. The MD averaged values and the corresponding standard deviations for the diameter are given in Table 1. Examining the standard deviations reveals that the fluctuation of the diameter increased upon complexation with MV in CB6 and CB7 but not in CB8. No noticeable change in the average value of the diameter upon inclusion of the guest molecule was observed for CB6 and CB8 (0.03 Å), but a change of 0.18 Å was measured for CB7. Free CB8 exhibits the largest fluctuation, while upon inclusion, CB7 fluctuates more.

***Scheme 1.*** Perpendicular Internal Diameters Traced in MD Simulation

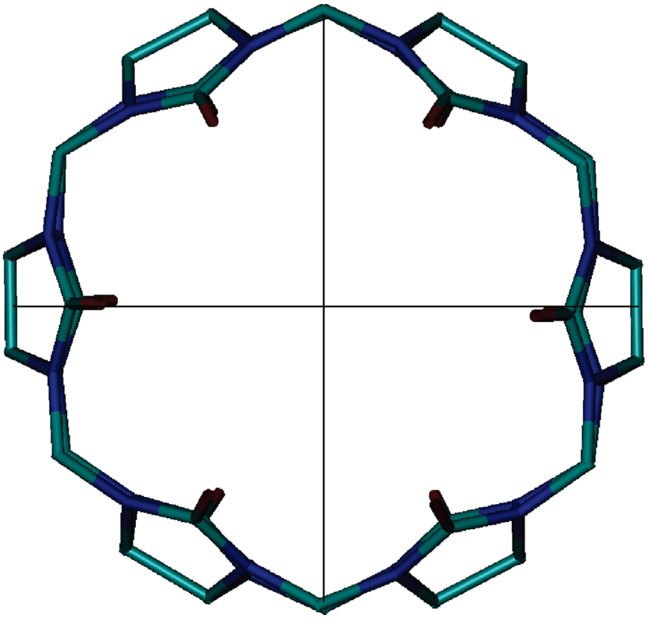


***Table 1.*** MD-Averaged Values of the Internal Diameter of CB

| complex | CB6 | CB6/MV | CB7 | CB7/MV | CB8 | CB8/MV |
|---|---|---|---|---|---|---|
| average diameter (Å) | 9.96 | 9.93 | 11.33 | 11.15 | 12.95 | 12.92 |
| standard deviation (Å) | 0.24 | 0.46 | 0.38 | 0.81 | 0.50 | 0.45 |

To better understand the conformational changes of the macrocyclic host upon guest inclusion, the values of two perpendicular internal diameters were traced in the MD simulation (Scheme 1). The absolute difference between the diameters as a function of simulation time and its distribution function for each system are given in Figure 5. Furthermore, the MD-averaged values and the corresponding standard deviations for the absolute difference are given in Table 2.

The difference between the diameters for CB6 shows a high probability around 0.2 Å for the free host. This value shifted to 0.8 Å in the complex. This indicates more sampling of a less symmetric structure than CB6 (oval shape) upon inclusion of the guest. The corresponding average value also increased after complexation (Table 2). This clearly demonstrates conformational changes of the macrocyclic structure induced by the guest inclusion. The situation is more dramatic in the case of CB7, upon whose inclusion, a severe broadening of the probability distribution peak was observed. The sampled structures for the complexed CB7 were found to span from 0 to 4 Å. The high extent of distortion in the case of CB7 (average value increased by ∼230%) may be explained by the complete inclusion of the guest molecule in CB7, compared to partial inclusion in CB6. Inspection of the CB7/MV trajectory indicates that the rotation of the molecule inside the cavity induces a change in the diameter that passes the guest molecular plane (Figure 6). Figure 5 shows that CB8 exhibits more or less similar distortion before and after complexation, although complete inclusion of the guest molecule was found in accord with the higher cavity size compared to MV size.

**Hydrogen Bond Analysis.** A summary of the intermolecular hydrogen bonds that exist between each CB and the surrounding water molecules is presented in Table 3. For hydrogen bond analysis, a hydrogen bond cut distance ≤ 3.0 Å and an angle ≥ 120° were used. Results in Table 3 demonstrate, as expected, that the number of hydrogen bonds increases as the number of glycouril units increases. The portal oxygen atoms in the glycouril unit establish hydrogen bonds with the water molecules nearby, while the nitrogen atoms in CB do not play a role in the hydrogen bond network with the solvent. All complexes showed a reduction in the total number of hydrogen bonds upon complexation, indicating rearrangement of water molecules near both rims in the inclusion process.

**MM/PBSA Results.** Table 4 lists the binding free energies (kcal/mol) resulting from the MM/PBSA analysis of the 5 ns MD trajectories obtained for the studied complexes. Results show that the major contribution to the binding free energy is the host−guest electrostatic interactions ($\Delta E_{\text{elec}}$), indicating the pronounced role of the ion−dipole interactions in such systems. CB7/MV shows the highest electrostatic contribution to the complex stability. This is expected since, unlike the CB6 complex, the two cationic groups interact with both carbonyl portals and associate in closer contact to the crowns than the CB8 complex does. van der Waals interactions ($\Delta E_{\text{vdW}}$) were also found to be the highest in the CB7 complex. This is a direct result of the complete inclusion of the hydrophobic part of the guest molecule into the CB7 cavity accompanied by the fact that the guest has a better complementary size to the CB7 cavity than the CB8 does. The internal bonded interaction ($\Delta E_{\text{INT}}$) values reflect significant conformational changes upon binding, especially in the CB6 system. The obtained $\Delta E_{\text{INT}}$ values are correlated well with the cavity size of the CB in which the values are getting more positive with decreasing the size of CB.

$\Delta G_{\text{NP}}$ values are negative for all complexes, indicating that the nonpolar surface term contributes positively to complex stability, though to a much lower extent when compared to the host−guest electrostatic and van der Waals interactions. Results also indicate unfavorable electrostatic solvation energy (positive values of $\Delta G_{\text{PB}}$) evidenced by an overall positive value of the solvation free energy for all complexes. This is attributed to the reduction of contact area with the solvent of both host and guest molecules upon inclusion and hence less electrostatic interactions of the complexed molecules with the solvent than the free molecules. The resulting rank of $\Delta G_{\text{PB}}$ (or $\Delta G_{\text{solv}}$) is expected, in which the CB7 complex exhibits the largest positive values, followed by CB8, and then CB6.

The computed binding free energies ($\Delta G$) reveal that the tendency of MV to complex with CBs is in the order CB7 > CB8 > CB6, in agreement with the experimental trend.

NMODE calculations indicate that negative values of $T\Delta S_{\text{conf}}$ are obtained for all studied complexes, thus demonstrating a loss in the degrees of freedom upon binding. The $T\Delta S_{\text{conf}}$ value obtained for the CB8 complex is more negative (−13.5 kcal/mol) when compared to the expected

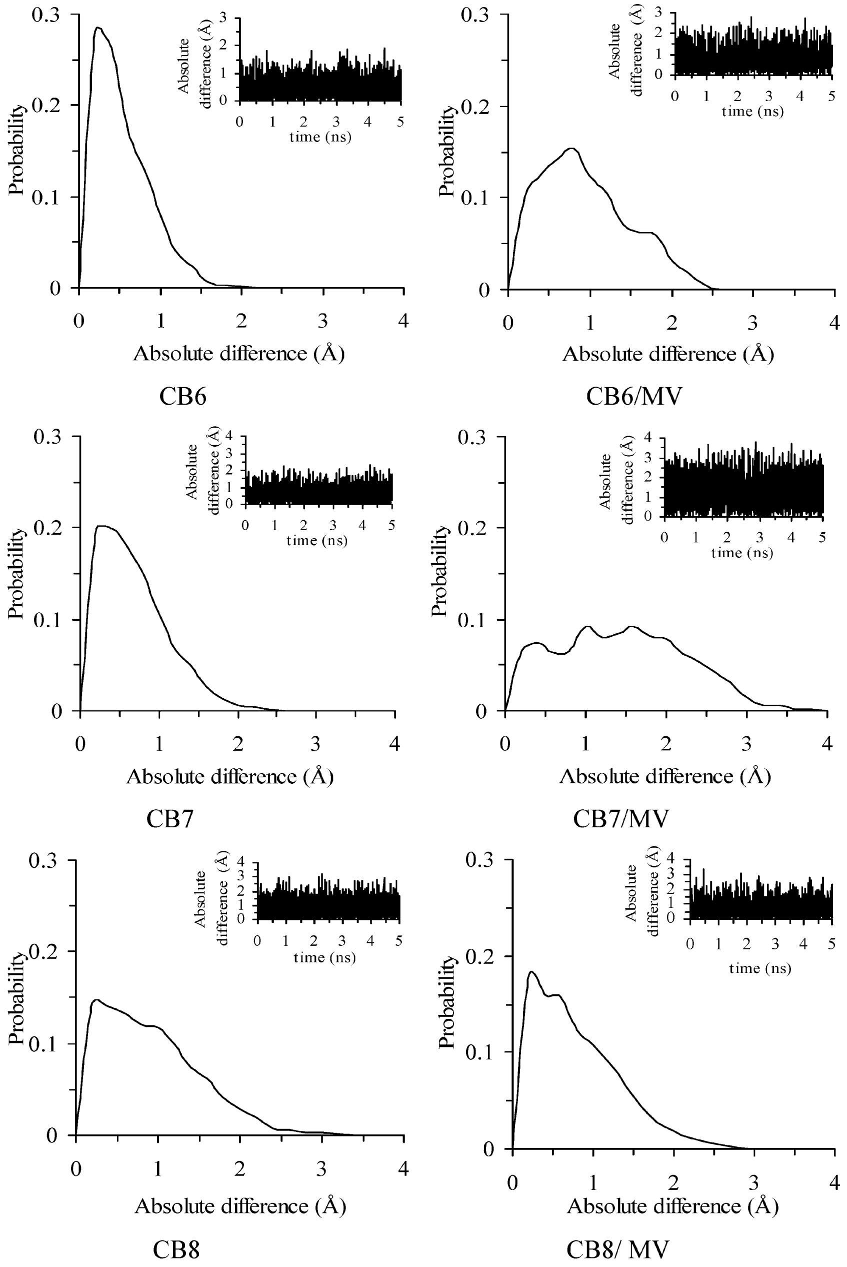


*Figure 5.* Distribution functions for the difference between two perpendicular internal diameters of CB. The difference as a function of simulation time is superimposed in each figure.

*Table 2.* MD-Averaged Values of the Absolute Difference between Two Perpendicular Internal Diameters of the Host

| complex | CB6 | CB6/MV | CB7 | CB7/MV | CB8 | CB8/MV |
|---|---|---|---|---|---|---|
| average difference (Å) | 0.44 | 0.85 | 0.59 | 1.36 | 0.82 | 0.72 |
| standard deviation (Å) | 0.34 | 0.54 | 0.43 | 0.78 | 0.58 | 0.54 |

stiffer CB7 complex (−11.5 kcal/mol). The results are in contrast with the fact that the guest molecule has more room in the CB8 cavity than CB7, thus having more mobility in CB8, as can be seen in Figure 3. This might be attributed to the approximate nature of NMODE calculations. However, results in Tables 1 and 2 and Figures 4 and 5 show that the fluctuation of the diameter and the difference between the perpendicular diameters increased upon complexation with MV in CB7 but slightly decreased in CB8. This might explain the obtained $T\Delta S_{\mathrm{conf}}$ values of CB7 and CB8 complexes.

It should be noted here that including the configurational entropy term in the binding free energy ($\Delta G^*$) also gives a similar trend of MV desiring to complex with CBs.

Attempts to conduct MD simulation of MV with $\beta$-cyclodextrin ($\beta$-CD) prove that MV does not enter into the $\beta$-CD cavity since the guest molecule escapes from the $\beta$-CD cavity after a few hundred picoseconds. This is due to the presence of two cationic groups in the guest molecule, which interacts more with the surrounding water molecules. This also indicates that, although there is a similarity in the sizes of the hydrophobic cavities in CB7 and $\beta$-CD, the interaction at the cavity entrances with guest molecule in both systems is different.[25] This result is in accord with the experimental

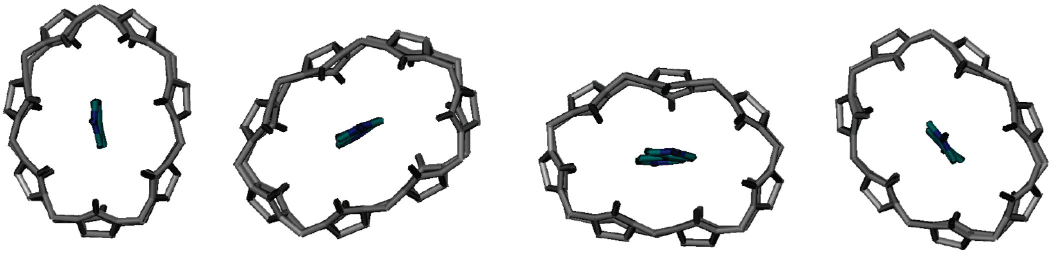


***Figure 6.*** Extracted snapshots of the CB7/MV complex to show the effect of guest rotation on the diameter of the host molecule.

***Table 3.*** Water−CB Intermolecular Hydrogen Bond Analysis for the Free and Complexed Host Molecules

| HB/complex | CB6 | CB6/MV | CB7 | CB7/MV | CB8 | CB8/MV |
|---|---|---|---|---|---|---|
| oxygen | 8.13 | 5.67 | 9.68 | 7.33 | 10.36 | 9.53 |
| nitrogen | | | 0.02 | | 0.01 | 0.01 |
| total | 8.13 | 5.67 | 9.70 | 7.33 | 10.37 | 9.54 |

***Table 4.*** Binding Free Energies (kcal/mol) Resulting from MM/PBSA Analysis of the Studied Complexes[a]

| | kcal/mol | CB6/MV | CB7/MV | CB8/MV |
|---|---|---|---|---|
| host | $E_{elec}$ | −1033.7 | −1118.3 | −1378.2 |
| | $E_{vdW}$ | −34.0 | −35.8 | −36.9 |
| | $E_{INT}$ | 257.6 | 303.1 | 350.7 |
| | $G_{NP}$ | 4.4 | 5.2 | 6.0 |
| | $G_{PB}$ | −135.1 | −167.7 | −181.0 |
| | $G_{solv}$ | −130.7 | −162.5 | −175.0 |
| | $TS$ | 81.7 | 94.5 | 107.2 |
| guest | $E_{elec}$ | | 142.7 | |
| | $E_{vdW}$ | | 12.7 | |
| | $E_{INT}$ | | 25.7 | |
| | $G_{NP}$ | | 2.1 | |
| | $G_{PB}$ | | −160.6 | |
| | $G_{solv}$ | | −158.5 | |
| | $TS$ | | 36.0 | |
| complex | $E_{elec}$ | −1003.1 | −1118.1 | −1360.0 |
| | $E_{vdW}$ | −46.5 | −54.8 | −47.9 |
| | $E_{INT}$ | 289.0 | 330.6 | 375.7 |
| | $G_{NP}$ | 4.6 | 4.9 | 6.0 |
| | $G_{PB}$ | −177.0 | −170.9 | −206.6 |
| | $G_{solv}$ | −172.4 | −166.0 | −200.6 |
| | $TS$ | 98.7 | 119.0 | 129.7 |
| $\Delta_{bind}$ | $-\Delta E_{elec}$ | −112.1 | −142.5 | −124.5 |
| | $\Delta E_{vdW}$ | −25.2 | −31.7 | −23.7 |
| | $\Delta E_{INT}$ | 5.7 | 1.8 | −0.7 |
| | $\Delta G_{NP}$ | −1.9 | −2.4 | −2.1 |
| | $\Delta G_{PB}$ | 118.7 | 157.4 | 135.0 |
| | $\Delta G_{solv}{}^{b}$ | 116.8 | 155.0 | 132.9 |
| | $\Delta G^{c}$ | −14.8 (2.2) | −17.4(2.1) | −16.0 (1.8) |
| | $T\Delta S$ | −19.0 (0.8) | −11.5 (0.7) | −13.5 (0.9) |
| | $\Delta G^{*d}$ | 4.2 (2.4) | −5.9 (2.2) | −2.5 (2.0) |
| | $\Delta G_{expt}$ | −1.8[e] | −7.2[f] | −6.1[g], −6.9[h] |

[a] Numbers in parentheses are standard deviations of the results. [b] $\Delta G_{solv} = \Delta G_{NP} + \Delta G_{PB}$ [c] $\Delta G = \Delta E_{elec} + \Delta E_{vdW} + \Delta E_{INT} + \Delta G_{NP} + \Delta G_{PB}$. [d] $\Delta G^{*} = \Delta G^{b} - T\Delta S_{conf}$ [e] Ref 24. [f] Ref 25. [g] Ref 27. [h] Ref 28.

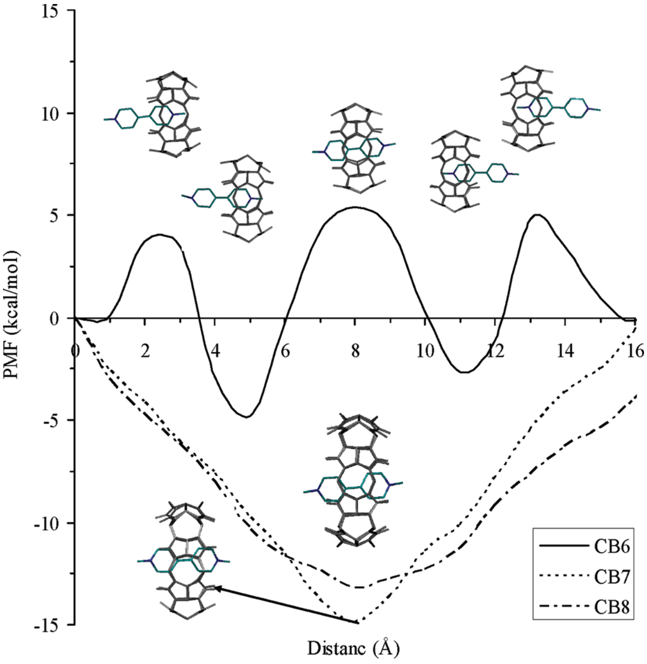


***Figure 7.*** PMF profiles for inclusion of MV.

data which showed that the $\beta$-CD/MV complex has a very low formation constant (∼0) with MV.[25] It is worth mentioning here the results of a recent MD study conducted by our group for the complexation of *N*-methyl-4-(*p*-methyl benzoyl)-pyridinium methyl cation with CB7 and $\beta$-CD. Results show that $\beta$-CD formed a more stable complex with the guest molecule than CB7 ($\Delta\Delta G \approx 1.9$ kcal/mol).[42] The guest molecule in that study possesses only one cationic site (methyl pyridinium cation), whereas MV has two cationic sites (two methyl pyridinium cations) capable of interacting with both CB7 portals by the favorable ion−dipole interactions, hence forming a more stable complex than $\beta$-CD. Moreover, the higher solubility of the dicationic guest compared to the monocationic guest is also responsible for a very small interaction with $\beta$-CD.

**Umbrella Sampling (PMF Calculations).** Figure 7 shows the results of PMF along the *r* coordinate (described in Figure 1). The PMF for the guest moving inside the CB6 cavity shows two minima at 5 and 11 Å with PMF values of −4.81 and −2.69 kcal/mol, respectively. Both minima correspond to more or less similar complex geometries which are composed of partial inclusion of the guest molecule. The difference in free energy between the two minima could be attributed to the distortion suffered by the macrocyclic compound when the guest passes from the unfavored complete inclusion to a partial inclusion, causing the difference in the free energy values. This is clear by the barrier existing at 8 Å which represents complete inclusion of the guest. The presence of the barrier explains the lack of sampling of snapshots composed of deep inclusion in the conventional MD simulation. There are also two barriers when the guest approaches both portals due to their small size. These barriers were not found for the bigger hosts, CB7 and CB8. The latter systems show only one minimum attributed to the complete inclusion geometry (8 Å).

## Conclusion

The complexation of methyl viologen with cucurbit[*n*]urils was studied by molecular dynamics simulations, MM-PBSA and umbrella sampling. Deep inclusion of the guest was observed in CB7 and CB8 complexes, while partial inclusion occurs in the case of CB6. The CB macrocycle distorted upon inclusion of the guest molecule, and this was clear when the dimensions of the cavity and the guest molecule were close. MM-PBSA calculations revealed that host−guest electrostatic interactions were of major importance for the stability of the complexes. The movement of the guest molecule through the CB6 cavity involved energy barriers, while no barriers were found in CB7 and CB8 for the guest movement.

**Acknowledgment.** The authors wish to thank the Hashemite University and Yarmouk University for their financial support. The authors also wish to thank Lutfi Hussein and Heather Hoyt for revising the language of the manuscript.

## References

(1) Rawashdeh, A. M.; Thangavel, A.; Sotiriou-Leventis, C.; Leventis, N. *Org. Lett.* **2008**, *10*, 1131.

(2) Lee, J. W.; Samal, S.; Selvapalam, N.; Kim, H.-J.; Kim, K. *Acc. Chem. Res.* **2003**, *36*, 621.

(3) Wang, R.; Macartney, D. H. *Tetrahedron Lett.* **2008**, *49*, 311.

(4) Lagona, J.; Mukhopadhyay, P.; Chakrabarti, S.; Isaac, L. *Angew. Chem., Int. Ed.* **2005**, *44*, 4844.

(5) Freeman, W. A.; Mock, W. L.; Shih, N. Y. *J. Am. Chem. Soc.* **1981**, *103*, 7367.

(6) Behrend, R.; Meyer, E.; Rusche, F. *Justus Liebigs Ann. Chem.* **1905**, *339*, 1.

(7) Kim, J.; Jung, I.-S.; Kim, S.-Y.; Lee, E.; Kang, J.-K.; Sakamoto, S.; Yamaguchi, K.; Kim, K. *J. Am. Chem. Soc.* **2000**, *122*, 540.

(8) Day, A. I.; Arnold, A. P.; Blanch, R. J.; Snushall, B. *J. Org. Chem.* **2001**, *66*, 8094.

(9) Day, A. I.; Blanch, R. J.; Arnold, A. P.; Lorenzo, S.; Lewis, G. R.; Dance, I. A. *Angew. Chem., Int. Ed.* **2002**, *41*, 275.

(10) Tuncel, D.; Steinke, J. H. G. *Chem. Commun.* **2001**, 253.

(11) Mock, W. L.; Pierpont, J. A. *J. Chem. Soc., Chem. Commun.* **1990**, 1509.

(12) Kornmüller, A.; Karcher, S.; Jekel, M. *Water Res.* **2001**, *35*, 3317.

(13) Jeon, Y. J.; Kim, S.-Y.; Ko, Y. H.; Sakamoto, S.; Yamaguchi, K.; Kim, K. *Org. Biomol. Chem.* **2005**, *3*, 2122.

(14) Wei, F.; Liu, S.-M.; Xu, L.; Cheng, G.-Z.; Wu, C.-T.; Feng, Y.-Q. *Electrophoresis* **2005**, *26*, 2214.

(15) Moch, W. L.; Shih, N. Y. *J. Org. Chem.* **1983**, *48*, 3618.

(16) Mezzina, E.; Cruciani, F.; Pedulli, G. F.; Lucarini, M. *Chem.*−*Eur. J.* **2007**, *13*, 7223.

(17) Choi, S. W.; Lee, J. W.; Ko, Y. H.; Kim, K. *Macromolecules* **2002**, *35*, 3526.

(18) Mock, W. L.; Shih, N. Y. *J. Am. Chem. Soc.* **1989**, *111*, 2697.

(19) Buschmann, H.-J.; Jansen, K.; Schollmeyer, E. *Acta Chim. Slov.* **1999**, *46*, 405.

(20) Li, Y.; Li, X.-Y.; Zhang, H.-Y.; Li, C.-J.; Ding, F. *J. Org. Chem.* **2007**, *72*, 3640.

(21) Sindelar, V.; Monn, K.; Kaifer, A. E. *Org. Lett.* **2004**, *6*, 2665.

(22) Sindelar, V.; Silvi, S.; Kaifer, A. E. *Chem. Commun.* **2006**, 2185.

(23) Zhang, H.; Paulsen, E. S.; Walker, K. A.; Krakowiak, K. E.; Dearden, D. V. *J. Am. Chem. Soc.* **2003**, *125*, 9284.

(24) Ong, W.; Gomez-Kaifer, M.; Kaifer, A. E. *Org. Lett.* **2002**, *4*, 1791.

(25) Kim, H.-J.; Jeon, W. S.; Ko, Y. H.; Kim, K. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 5007.

(26) Moon, K.; Kaifer, A. E. *Org. Lett.* **2004**, *6*, 185.

(27) Xiao, X.; Tao, Z.; Xue, S.-F.; Zhu, Q.-J.; Zhang, J.-X.; Lawrance, G. A.; Raguse, B.; Wei, G. *J. Inclusion Phenom. Macrocyclic Chem.* **2008**, *61*, 131.

(28) Jeon, W. S.; Kim, H.-J.; Lee, C.; Kim, K. *Chem. Commun.* **2002**, 1828.

(29) Frisch, M.; Trucks, G.; Schlegel, H.; Scuseria, G.; Robb, M.; Cheeseman, J.; Montagomery, J.; Verven, J. R.; Kudin, K.; Burant, J.; Millam, J. M.; Iyenger, S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G.; 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.; Konx, J.; Hratchian, H.; Cross, J.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R.; Yazyev, O.; Austin, A.; Cammi. R.; Pomelli, C.; Ochterski, J.; Ayala, P.; Moromuka, K.; Voth, G.; Salvador, P.; Dannenberg, J.; Zakrzewski, V.; Dapprich, S.; Daniels, A.; Strain, M.; Farkas, O.; Malick, D.; Rabuck, A.; Raghavachari, K.; Foresman, J.; Ortiz, J.; Cui, Q.; Baboul, A.; Clifford, S.; Cioslwski, J.; Setvanov, B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Marton, R.; Fox, D.; Keith, T.; Al-Laham, M.; Peng, C.; Nanayakkara, A.; Challacombe, M.; Gill, P.; Johnson, B.; Chen, W.; Wong, M.; Gonzalez, C.; Pople, J. *Gaussian 03*, Revision D.01; Gaussian, Inc.: Wallingford, CT, 2004.

(30) Bayly, C.; Cieplak, P.; Cornell, W.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269.

(31) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. *J. Comput. Chem.* **2000**, *21*, 132.

(32) Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R. L.; 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. *AMBER8*; University of California: San Francisco, CA, 2004.

(33) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157.

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

(35) Darden, T.; York, D.; Pederson, L. *J. Chem. Phys.* **1993**, *98*, 10089.

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

(37) Sanner, M. F.; Olson, A. J.; Spehner, J. C. *Biopolymers* **1996**, *38*, 305.

(38) Hou, T. J.; Guo, S. L.; Xu, X. J. *J. Phys. Chem. B* **2002**, *106*, 5527.

(39) El-Barghouthi, M. I.; Jaime, C.; Al-Sakhen, N. A.; Issa, A. A.; Abdoh, A. A.; Al-Omari, M. M.; Badwan, A. A.; Zughul, M. B. *THEOCHEM* **2008**, *853*, 45.

(40) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. *J. Comput. Chem.* **1992**, *13*, 1011.

(41) Souaille, M.; Roux, B. *Comput. Phys. Commun.* **2001**, *135*, 40.

(42) Rawashdeh, A. M.; El-Barghouthi, M. I.; Assaf, K. I.; Al-Gharabli, S. I. *J. Inclusion Phenom. Macrocyclic Chem.* **2009**, *64*, 357.

CT900622H

JCTC Journal of Chemical Theory and Computation

# Theoretical Study of the Hydrolysis of Pentameric Aluminum Complexes

Jaakko Saukkoriipi[†] and Kari Laasonen*

*Department of Chemistry, University of Oulu, P.O. Box 3000, Oulu FI-90014, Finland*

Received December 14, 2009

**Abstract:** Static quantum chemical calculations and Car−Parrinello molecular dynamics (CPMD) simulations were used to investigate the structural characteristics and the stability of pentameric aluminum clusters in both gas phase and aquatic environments. The accuracy of several generalized gradient approximation (GGA) and hybrid exchange-correlation functionals were tested with and without empirical van der Waals corrections to ensure the accuracy of the selected methods. Conformational analysis was performed for experimentally detected (electrospray ionization mass spectrometry, ESI MS) structural isomers of cationic $[Al_5O_6H_2Cl_4]^+$, $[Al_5O_7H_4Cl_4]^+$, $[Al_5O_8H_6Cl_4]^+$, and $[Al_5O_9H_8Cl_4]^+$ complexes. Conductor-like screening model (COSMO) was used to investigate the stability of the gas phase optimized structures in aquatic environments. Four of the main pentameric aluminum complexes were then selected for the CPMD investigation. The effect of the long-range empirical vdW corrections (-D) was also tested employing two identical simulations, with and without the corrections. During these simulations, several spontaneous associative hydration reactions were detected. The open and highly symmetric hexagonal prism-like structures were found to be dominating geometries in liquid conditions. Overall, CPMD calculations produced distinctly different geometries for the pentameric molecules than for the static calculations.

## 1. Introduction

The speciation of aluminum compounds in aquatic environments has been widely studied both experimentally and theoretically during the past few decades. Experimental methods, such as potentiometric titration,[1,2] $^{27}$Al nuclear magnetic resonance (NMR) spectroscopy,[3−10] X-ray crystallography,[11,12] and electrospray ionization mass spectrometry (ESI MS)[13−17] have been used to reveal the hydrolysis of aluminum salts during coagulation in aquatic solutions. At the same time, several computational studies have addressed the structures and energetics of different sized aluminum compounds and ions in gas phase and in liquid environments.[18−29] As a result of these investigations, a wide variety of different aluminum species have been detected from small monomeric to large polynuclear aluminum complexes. However, there are still a number of unknown factors concerning the exact nature and composition of the hydrolysis products of aluminum salts in aqueous environments. One of the most interesting factors is the role of small oligomeric species, like pentameric aluminum complexes in the aforementioned hydrolysis processes.

The first real evidence of the existence of pentameric aluminum complexes in aquatic solutions was detected in the high-field $^1$H and $^{27}$Al NMR studies of Akitt et al.[30] They postulated that, at higher aluminum concentrations (>0.02 mol dm$^{-3}$) and at intermediate levels of hydrolysis, an oligomeric mixture of $[Al(OH)_{2.5}]^{0.5+}$ is formed, which can then be composed of pentameric aluminum complexes, like $[Al_5(OH)_{12}]^{3+}$ and $[Al_5(OH)_{13}]^{2+}$.[30] However, the results of Akitt et al. indicated that these oligomeric aluminum species were only intermediate forms in polymerization from monomers to the tridecameric “Keggin” cation $[AlO_4Al_{12}(OH)_{24}(H_2O)_{12}]^{7+}$, especially in low aluminum concentrations

* Corresponding author phone: +358 8 553 1640; fax: +358 8 553 1603; e-mail: kari.laasonen@oulu.fi.

[†] Present address: Finnish Environment Institute (SYKE), Freshwater Centre, River Basin Management, The Oulu Office, P.O. Box 413, FI-90014, University of Oulu, Finland. E-mail: jaakko.saukkoriipi@ymparisto.fi.

10.1021/ct900670a © 2010 American Chemical Society
Published on Web 03/22/2010

(<0.02 mol dm$^3$).[30,31] Two decades later, Sarpola et al. investigated the hydrolysis products of aluminum in the aquatic environments using the ESI MS method and detected around 20 different cationic pentameric aluminum complexes mainly at 10−100 mM aluminum concentrations in acidic solutions (pH < 4.7).[13−15] They also postulated that pentameric structures seemed to be surprisingly stable in aquatic environments.[32]

Recently, Zhao et al. studied the effect of pH to the hydrolysis of aluminum salts using the ESI MS method and discovered only one pentameric aluminum complex ($[Al_5O_7]^+$), which was present mainly at the pH range of 4.0−5.0.[17] Zhao et al. postulated also two mechanisms for the coagulation of aluminum salts, the “core-link” and “cage-like” models. In the “cage-like” model, there are only monomeric, dimeric, and tridecameric polycations and larger polynuclear aluminum species in the aquatic solutions.[17] However, the “core-links” model gives a distribution of continuously changed aluminum compounds from monomeric species to smaller oligomeric aluminum complexes, etc.[17] Zhao et al. claimed that pentameric aluminum complexes can either fragment to tetrameric aluminum compounds by loosing an aluminum atom or aggregate by self-assembly to form larger aluminum complexes ($Al_{10}$) in aquatic environments.[17] They also postulated that small oligomeric aluminum compounds ($Al_3$−$Al_5$) are dominating in the pH range of 4.6−4.8 at very low aluminum concentrations ($1.5 \times 10^{-4}$ mol dm$^{-3}$).

Das et al. combined quantum chemical density functional calculations with anion photoelectron and mass spectroscopy. They investigated the structural characteristics of neutral and anionic $Al_5O_4$ clusters.[33,34] Highly symmetric planar geometry was detected to be the ground state structure for the anionic aluminum oxide cluster $Al_5O_4^-$. In addition, a neutral $Al_5O_4$ cluster was observed to have very large electron affinity, equivalent with the affinity of the chlorine ion.[34] The results of the photoelectron spectroscopy revealed the high reactivity of the anionic cluster toward an aqua ligand. Reactivity was then confirmed in the static quantum chemical density functional calculations.[33] Although pentameric aluminum clusters have been studied both experimentally and theoretically, there are still a number of uncertainties concerning their stability and structural characteristics in the hydrolysis processes.

In this investigation, we have studied the characteristics of cationic pentameric aluminum complexes in aquatic environments using static quantum chemical methods and Car−Parrinello molecular dynamics (CPMD) simulations. Pentameric complexes were taken directly from the ESI MS results of Sarpola et al.[13−15,32] It should be noted that most of the time, especially when the fragmentation series of aqua ligands or isotopic patterns are lacking in the spectra, ESI MS gives only the sum mass of the complex.[14] Hence, quantum chemical conformational analysis was performed to obtain the lowest energy conformations of the chosen pentameric species prior to the ab initio molecular dynamics (AIMD) simulations. The main goal of this study was to reveal the stability of the pentameric aluminum complexes in aquatic environments and, in addition, to elucidate their role in the aforementioned hydrolysis processes. The secondary objective was to investigate the applicability of the empirical van der Waals corrections to both static calculations and CPMD simulations.

## 2. Computational Details

**2.1. Static, Gas Phase Studies.** We investigated the structures of aluminum complexes with the molecular formulas of $[Al_5O_6H_2Cl_4]^+$, $[Al_5O_7H_4Cl_4]^+$, $[Al_5O_8H_6Cl_4]^+$, and $[Al_5O_9H_8Cl_4]^+$. We note that these were the main pentameric aluminum species detected in the ESI MS studies of Sarpola et al.[13−15,32] Geometry optimizations were carried out without symmetry constraints. The structural optimization of the clusters were performed using density functional theory (DFT) with the Perdew−Burke−Ernzerhof (PBE) functional[35] and polarized valence triple ζ (TZVP) basis set.[36] Calculations were performed using the Turbomole 6.0 program suite.[37] Resolution-of-the-identity (RI) approximation was used to accelerate the calculations.[38]

The choice of PBE density functional was justified with a test of four trimeric, four tetrameric, and four pentameric aluminum (chloro)hydroxide complexes, which were then optimized in the gas phase with different methods (Becke88 exchange and Lee−Yang−Parr correlation functional (BLYP), Becke’s three parameter hybrid functional (B3LYP)) with the TZVP basis set. The accuracy and the effect of empirical van der Waals corrections (-D) were also tested here (PBE and B3LYP).[39] For the usage of these corrections, three sets of parameters must be defined, density functional dependent global scaling factor $s_6$, dispersion coefficients $C_6$ ([Jnm$^6$ mol$^{-1}$]), and van der Waals radii $R_0$ ([Å]) for the elements.[40] Scaling factor $s_6$ for PBE was 0.75 and 1.05 for B3LYP. The $C_6$ and $R_0$ parameters for aluminum were 10.79 and 1.639, for oxygen 0.70 and 1.342, for hydrogen 0.14 and 1.001, and for chlorine 5.07 and 1.639. The total energy is then given by equation

$$E_{\mathrm{DFT-D}} = E_{\mathrm{DFT}} + \left(-s_6 \sum_{i=1}^{N_{\mathrm{at}}-1} \sum_{j=i+1}^{N_{\mathrm{at}}} \frac{C_6^{\;ij}}{R_{ij}^{\;6}} f_{\mathrm{dmp}}(R_{ij})\right) \qquad (1)$$

$N_{\mathrm{at}}$ denotes the number of atoms in a system, $C_6^{ij}$ is the dispersion coefficient for atom pair $ij$ that can be written as geometric mean $((C_6^{\,i}C_6^{\,j})^{1/2})$. We note that the selected trimeric and tetrameric complexes came from the previous studies of the author.[41] Second-order Møller−Plesset perturbation theory (MP2) with the frozen core approximation and quadruple ζ valence with double polarization (QZVPP) was chosen for the reference method, see Table 1.[42−44]

The justification of the frozen core approximation is that the inner-shell electrons of an atom are less sensitive to their environment than the valence electrons. Thus, the error introduced by freezing the core orbitals is nearly constant for molecules containing the same types of atoms. The accuracy of this approximation was also tested by fully optimizing the selected aluminum structures in the MP2/QZVPP level of theory with and without frozen core approximation. Approximation clearly accelerated the optimization procedure but had only a mild affect to the relative

***Table 1.*** Comparison of the Gas Phase Energy Differences [kJ mol$^{-1}$] of Aluminum Chlorohydrates

| cluster | PBE | BLYP | B3LYP[a] | PBE-D | B3LYP-D[a] | MP2/QZVPP |
|---|---|---|---|---|---|---|
| $[Al_3O_7H_9Cl_3]^+$_1 | 19.0 | 25.3 | 25.5 | 26.5 | 35.7 | 23.3 |
| $[Al_3O_7H_9Cl_3]^+$_2 | 32.9 | 27.0 | 28.6 | 32.9 | 28.5 | 32.7 |
| $[Al_3O_7H_9Cl_3]^+$_3 | 3.1 | 1.5 | 2.0 | 3.3 | 1.9 | 4.2 |
| $[Al_3O_7H_9Cl_3]^+$_4 | 0 | 0 | 0 | 0 | 0 | 0 |
| $[Al_4O_8H_9Cl_4]^+$_1 | 25.7 | 22.5 | 23.5 | 27.8 | 25.9 | 24.8 |
| $[Al_4O_8H_9Cl_4]^+$_2 | 57.4 | 47.5 | 51.2 | 49.8 | 40.2 | 52.3 |
| $[Al_4O_8H_9Cl_4]^+$_3 | 35.0 | 24.1 | 32.7 | 39.4 | 33.7 | 45.5 |
| $[Al_4O_8H_9Cl_4]^+$_4 | 0 | 0 | 0 | 0 | 0 | 0 |
| $[Al_5O_7H_4Cl_4]^+$_1 | 56.9 | 44.8 | 50.0 | 45.9 | 34.4 | 49.3 |
| $[Al_5O_7H_4Cl_4]^+$_2 | 40.4 | 28.3 | 31.9 | 35.6 | 25.3 | 38.2 |
| $[Al_5O_7H_4Cl_4]^+$_3 | 49.8 | 53.0 | 55.9 | 47.4 | 53.0 | 49.9 |
| $[Al_5O_7H_4Cl_4]^+$_4 | 0 | 0 | 0 | 0 | 0 | 0 |
| average[b] | 2.7 | 4.7 | 3.1 | 2.0 | 6.2 | |
| standard deviation[b] | 3.5 | 6.0 | 3.8 | 1.9 | 6.0 | |

[a] Without resolution-of-the-identity (RI) approximation. [b] Calculated from the absolute values of $|\Delta E_{MP2} - \Delta E_{DFT}|$.

energy differences between chosen aluminum conformations (0.1−5.9 kJ mol$^{-1}$). The accuracy of the chosen theory was also tested against the second-order approximate coupled-cluster (CC2/QZVPP) method. This was done optimizing five dimeric ($[Al_2O_6H_9Cl_2]^{+}$ [41]), and eight trimeric (four $[Al_3O_7H_8Cl_3]^+$ and four $[Al_3O_7H_8SO_4]^+$)[16,41] aluminum conformations with the CC2 level of theory without frozen core approximation and comparing the relative energy differences to the MP2 results. The linear regression analysis of the relative energy differences were very close to linear dependency; the goodness of fit ($R^2$) was 0.9996. The results of these tests confirmed clearly that the accuracy of the MP2/QZVPP/frozen core approximation level of theory is sufficient for the reference method.

The comparison of the accuracy of the chosen methods (Table 1) was performed by subtracting the relative energy differences of trimeric, tetrameric, and pentameric aluminum complexes calculated with different density functional or hybrid methods from the corresponding relative energy differences of the reference method. This was followed by taking the absolute values of the extractions. Then, the averages and standard deviations were calculated of the absolute values; see Table 1. Note that the ground state energies were also included ($n = 12$). The results showed that the PBE functional gives slightly more consistent results compared to the BLYP density functional[45,46] and B3LYP hybrid functional[45−48] with lower mean value and standard deviation. The van der Waals-corrected functionals yielded the following results; the PBE-D had the lowest mean value and standard deviation of the test, whereas B3LYP-D had the highest. The results indicated clearly that the empirical corrections can either enhance the accuracy of the functional (PBE) or worsen it (B3LYP). Due to this test, the PBE density functional with triple $\zeta$ basis set was chosen for the preliminary optimizations and the PBE with empirical van der Waals corrections was chosen for the verification.

During the test, we examined also the double-hybrid functional B2PLYP-D with long-range empirical dispersion corrections.[45−47,49,50] We tested its accuracy using the second-order approximate coupled cluster (CC2/QZVPP) as the reference, although B2PLYP-D can only be used for single point calculations in Turbomole. It is a new type of hybrid density functional with global parameters of $\mathbf{a_x}$ for describing the mixture of Hartree−Fock (HF) and generalized gradient approximation (GGA) exchange and **c** for describing the perturbative second-order correlation part (PT2) and GGA correlation.[49,50] The relative energy differences of different dimeric and trimeric aluminum chlorohydrate conformations calculated with the B2PLYP-D/TZVP level of theory were then subtracted from CC2/QZVPP results taking the absolute values from the differences. The high standard deviation (15.6) and rather high mean value (8.2) of the differences indicate that the new double-hybrid density functional (DHDF) is not the best choice for describing the chemical nature and relative energy differences of aluminum (chloro)hydroxide complexes. The new DHDF was also tested for tetrameric and pentameric aluminum chlorohydrates using MP2/QZVPP with frozen core approximation as a reference. In the case of small oligomeric aluminum complexes, not only the standard deviations and mean values increased but also the functional failed to describe the trends in energy differences of different conformations compared to the results of any other density functional or reference method. These observations show clearly that the B2PLYP-D is not a suitable method for cationic aluminum (chloro)hydroxide calculations and, hence, was not used in further investigations in this study.

**2.2. Static, Liquid Phase Calculations.** The stability of gas phase optimized structures in aqueous environments was investigated using conductor-like screening model (COSMO) with triple $\zeta$ valence with double polarization.[36,51,52] COSMO is a solvation model, where the solute forms a cavity within the dielectric continuum of the permittivity $\varepsilon$ that represents the solvent.[37] In this study, water ($\varepsilon = 78.39$) was chosen for the solvent. Most of the parameters employed were the default parameters of COSMO, e.g., optimized van der Waals radii for O, H, and Cl atoms existed in the code. The radius for chlorine was 2.05 Å, for oxygen 1.72 Å, and for hydrogen 1.30 Å, and the scaling factor was approximately 1.89. However, the van der Waals radius of the aluminum ion ($R_{Al}$) had to be defined computationally. It was determined as follows. The literature value for the Gibbs free energy of hydration of $Al^{3+}$ ion is −4619 kJ mol$^{-1}$.[53] Burgess introduced Gibbs free energies of the hydration for the cations relative to the estimated free energy of the hydration of a proton (−1090.7 kJ mol$^{-1}$).[53] Tissandier et al. corrected and updated the value of the absolute Gibbs free energy of the hydration of the proton by less than 14 kJ mol$^{-1}$ (−1104.5 kJ mol$^{-1}$) in their Cluster-Pair-Based approximation studies.[54] The value of the Gibbs free energy of aluminum ion was then corrected according to the correction of Tissandier et al.[54]

The optimized COSMO radius for aluminum ion was calculated using eq 2, where $\Delta E_{\mathrm{Cosmo}}(\mathrm{Al}^{3+},R_{\mathrm{Al}})$ is the COSMO-corrected total energy of $\mathrm{Al}^{3+}$ ion and $\Delta E_{\mathrm{Vacuum}}(\mathrm{Al}^{3+})$ is the total energy of $\mathrm{Al}^{3+}$ ion in gas phase. The optimized van der Waals radius for aluminum ion was then specified to 1.3287 Å. Coskuner et al. postulated a similar van der Waals radius for aluminum (1.33) in their coordination studies of Al-EDTA in aqueous solutions.[55]

$$\Delta E_{\mathrm{Solv}}(\mathrm{Al}^{3+}) = \Delta E_{\mathrm{Cosmo}}(\mathrm{Al}^{3+},R_{\mathrm{Al}}) - \Delta E_{\mathrm{Vacuum}}(\mathrm{Al}^{3+}) \qquad (2)$$

During this research, all COSMO calculations were made as single point calculations. The choice was made to compare the solvation energy differences of gas phase optimized conformations without altering the gas phase structures. The choice can be rationalized with the fact that the measuring process in the ESI MS method takes place in vacuum conditions.[13−16] However, the choice of single point COSMO calculations has been tested in reference to the full COSMO optimization calculations in our previous computational studies.[16,41] More detailed description of the tests can be found in refs 16 and 41.

**2.3. Car−Parrinello Molecular Dynamics Simulations.** Pentameric aluminum complexes were also studied in a liquid environment using ab initio molecular dynamics (AIMD)-simulations. The selected pentameric species were $[\mathrm{Al_5O_6H_2Cl_4}]^+$ ($m/z$ = 373) and $[\mathrm{Al_5O_7H_4Cl_4}]^+$ ($m/z$ = 391). The initial molecular formulas for the pentameric aluminum clusters came from the ESI MS studies of Sarpola et al.[13−16,32] The minimum energy structures for chosen oligomeric aluminum clusters were then deduced in gas phase in the static part of this study. Three different conformations including the gas phase minima of the cationic $[\mathrm{Al_5O_6H_2Cl_4}]^+$ complex were taken and placed into a 17.0 Å cubic shell and solvated by 141 explicit water molecules, producing a density of 1.04 g cm$^{-3}$. In addition, the gas phase minima of the cationic $[\mathrm{Al_5O_7H_4Cl_4}]^+$ cluster was solvated by 169 water molecules in a cubic shell of 18.0 Å sides, producing a density of 1.04 g cm$^{-3}$. Note that densities were calculated for a deuterated system. The initial guess of the water molecule positions was based on the simple point charge (SPC) water model.[56,57]

We used 24 Ry cutoffs for the plane wave expansion and periodic boundary conditions. Simulations were performed in the canonical ensembles (NVT) using a Car−Parrinello molecular dynamics approach (CPMD).[58] The temperature of the simulations was scaled to 350 K using a chain of Nose-Hoover thermostats[59−62] with characteristic frequency of 2500 cm$^{-1}$, which was fixed high enough to ensure that the OH moieties were thermostatted properly and to have better control of fictitious kinetic energy. The core electrons were described using Vanderbilt ultra soft pseudopotentials[63−65] for all atoms in a system. The time step for electrons and ions in these simulations was 6 atomic units (0.145 fs), which was possible using deuterium instead of hydrogen in the simulations. The following atomic masses for the nuclei were used 2.0 amu for hydrogen, 16.0 amu for oxygen, 27.0 amu for aluminum, and 35.4 amu for chlorine atoms. Fictitious electron mass ($\mu$) was 650 atomic units, and the total charge of the system was +1. The equations of motion were solved using velocity Verlet algorithm.

PBE density functional[35] was used throughout the simulations. It was selected for the investigations due to the proven accuracy in describing the structural characteristics and stability of the aluminum chlorohydroxides in aquatic environments.[66] The effect of empirical van der Waals corrections was also tested during this research. Simulations were performed with and without van der Waals corrections. Corrections were employed using the ALL GRIMME approach, where long-range dispersion forces are considered by explicitly including damped pairwise interatomic potentials of $C_6^{ij}R_{ij}^{-6}$ form in the total energy.[39,40] The average simulation times were from 30 to 45 ps throughout this investigation.

## 3. Results and Discussion

**3.1. Static Calculations.** In this part, we will focus on the structural characteristic of the cationic pentameric aluminum complexes, $[\mathrm{Al_5O_6H_2Cl_4}]^+$, $[\mathrm{Al_5O_7H_4Cl_4}]^+$, $[\mathrm{Al_5O_8H_6Cl_4}]^+$, and $[\mathrm{Al_5O_9H_8Cl_4}]^+$. We will discuss the conformational isomers of the aluminum clusters above and compare the characteristics of the most interesting low energy conformations. In total, we have investigated hundreds of planar and nonplanar configurations of the oligomeric aluminum species for the geometry optimizations. The structures were optimized in gas phase using the PBE/TZVP approach. The optimization procedure was then divided into two stages: first, the core of the clusters was optimized, followed by the optimization of the ligand orientations. The lowest energy conformations including the gas phase minima were then selected for the cross-checking with PBE-D/TZVPP level of theory and for the COSMO calculations in order to compare their relative energy differences in aquatic solutions.

We first focus on the structural characteristics and solvation of the cationic $[\mathrm{Al_5O_6H_2Cl_4}]^+$ ($m/z$ = 373) aluminum (chloro)hydroxide clusters. The core of the gas phase minimum of this cluster is composed of a netlike structure of $\mathrm{Al_5O_4}$ and resembles closely the minimum energy structure of $[\mathrm{Al_5O_4}]^-$ anion introduced in the studies of Das et al.[33,34] However, the two hydroxo bridges between the corner aluminum atoms are bending the core structure from the symmetric planar configuration to a asymmetric form, see structure (**a**) in Figure 1. The minimum energy structure seems to be a combination of a trimeric ($[\mathrm{Al_3O_3(OH)Cl_2}]^0$) with $\mathrm{C_{3v}}$ symmetry and dimeric ($[\mathrm{Al_2O(OH)Cl_2}]^+$) aluminum complexes, which both have cores resembling closely to the dimeric and trimeric aluminum (chloro)hydroxides introduced in our previous studies.[41,66] However, these substructures were not seen individually in the ESI MS studies of Sarpola et al.[13−16]

The structure (**a**) consists of four equivalent three coordinated oxygen atoms and two hydroxo bridges. The core aluminum−oxygen ($R_{\mathrm{Al-O}}$) bond distances varied from 1.79

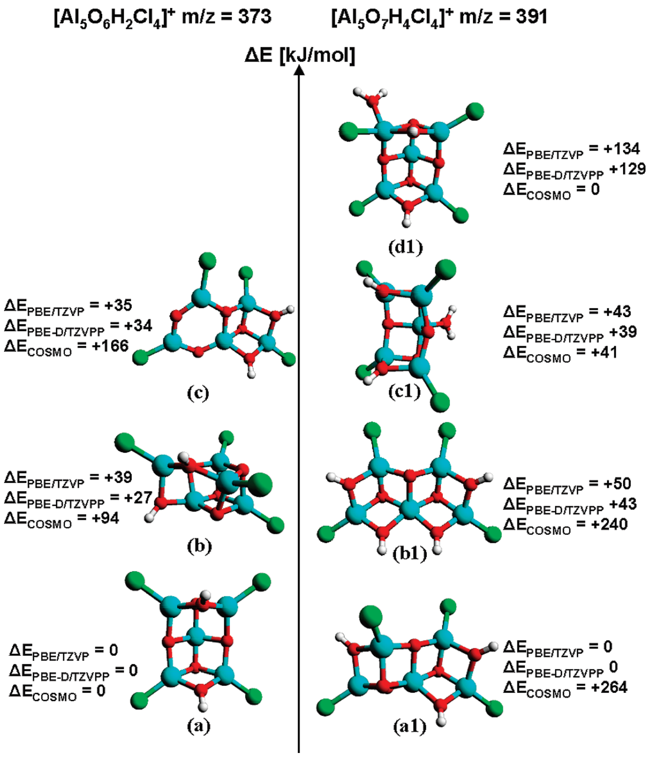

$[Al_5O_6H_2Cl_4]^+$ m/z = 373
$[Al_5O_7H_4Cl_4]^+$ m/z = 391
ΔE [kJ/mol]
$\Delta E_{PBE/TZVP}$ = +134
$\Delta E_{PBE-D/TZVPP}$ +129
$\Delta E_{COSMO}$ = 0
(d1)
$\Delta E_{PBE/TZVP}$ = +35
$\Delta E_{PBE-D/TZVPP}$ = +34
$\Delta E_{COSMO}$ = +166
(c)
$\Delta E_{PBE/TZVP}$ = +43
$\Delta E_{PBE-D/TZVPP}$ +39
$\Delta E_{COSMO}$ = +41
(c1)
$\Delta E_{PBE/TZVP}$ = +39
$\Delta E_{PBE-D/TZVPP}$ = +27
$\Delta E_{COSMO}$ = +94
(b)
$\Delta E_{PBE/TZVP}$ = +50
$\Delta E_{PBE-D/TZVPP}$ +43
$\Delta E_{COSMO}$ = +240
(b1)
$\Delta E_{PBE/TZVP}$ = 0
$\Delta E_{PBE-D/TZVPP}$ = 0
$\Delta E_{COSMO}$ = 0
(a)
$\Delta E_{PBE/TZVP}$ = 0
$\Delta E_{PBE-D/TZVPP}$ 0
$\Delta E_{COSMO}$ = +264
(a1)

***Figure 1.*** Optimized low energy configurations for the cationic $[Al_5O_6H_2Cl_4]^+$ and $[Al_5O_7H_4Cl_4]^+$ aluminum complexes. The energy differences are calculated relative to the gas phase and COSMO minima. Aluminum is presented in blue, oxygen in red, chlorine in green, and hydrogen in white.

to 1.89 Å, being the shortest between the center aluminum atom and trivalent oxygen bridges located between corner aluminum atoms. The corner aluminum and hydroxo bridge oxygen atom ($R_{Al-OH}$) bond lengths were around 1.86 Å, whereas the corner aluminum atom and trivalent core oxygen atom bond distances varied from 1.83 Å to 1.89 Å, indicating the asymmetry of the structure. All aluminum−oxygen bond lengths mentioned above differed slightly from the equivalent bond distances in the $[Al_5O_4]^-$ anion of Das et al.[33,34] This is mainly due to the strongly electronegative chlorido ligands attached to all four corner aluminum atoms. The chlorido ligands attract valence electrons from the aluminum atoms, thereby weakening other bonds of the cluster. The aluminum-chlorine ($R_{Al-Cl}$) bond lengths were around 2.07 Å. Despite this ligand effect, all aluminum−oxygen ($R_{Al-O}$) bond lengths were within the typical Al−O single bond distance range.[67,68]

According to the previous studies, aluminum prefers octahedral coordination in aquatic environments.[27,28,69] However, also 5-fold coordination has been detected for the aluminum ions in solutions.[66,70] In the case of cationic $[Al_5O_6H_2Cl_4]^+$ minimum, all five aluminum atoms in the structure were four coordinated, indicating that the complex is an ideal Lewis acid for accepting an electron pair from surrounding aqua ligands in the aquatic environments. In other words, it has a vacant coordination position in the valence shell making it ideal for acting as an acceptor of a new donor aqua ligand. Due to these hypotheses, the gas phase minima of this oligomeric aluminum complex was chosen for the CPMD simulations.

The structure (**b**) was chosen here not only due to its low energy but also because it has a core structure that resembles closely to the core of an intermediate tetrameric aluminum chlorohydrate $[Al_4O_2(OH)_3Cl_4]^+$ form between adamantane-like and cubane-like structures.[41] The aluminum−oxygen ($R_{Al-O}$) bond lengths varied from 1.78 to 1.89 Å, and the $R_{Al-Cl}$ bonds varied from 2.05 to 2.08 Å, respectively. The structure had also four equivalent trivalent oxygen atoms bridging simultaneously three different aluminum atoms leading to a tetrahedral coordination of the aluminum atoms. Structure (**b**) was also selected for the CPMD investigations in order to investigate its stability in aquatic environments.

The structure (**c**) can be considered as a combination of neutral dimeric $[Al_2O_2Cl_2]^0$ and cationic trimeric $[Al_3O_2(OH)_2Cl_2]^+$ ($C_{3v}$ symmetry) aluminum complexes, as seen in Figure 1. The core structure has a hexagon part and a compact trimeric part sharing one aluminum and one oxygen atom. It consists of two trivalent oxygen, two bivalent oxygen atoms, and two hydroxo bridges. The aluminum−oxygen bond distances varied from 1.69 to 1.88 Å, being the shortest on the aluminum and oxygen atoms in the hexagon. The aluminum−chlorine bond distances were around 2.07 Å. Structure (**c**) differs from the rest of the conformations by having two three coordinated aluminum atoms in the hexagon side. However, all three aluminum atoms in the trimeric side preferred 4-fold coordination. The structure (**c**) was also selected for the CPMD simulations. The results of the simulations will be discussed more closely in the following sections.

The stability of the gas phase optimized structures was then investigated in aquatic environments with COSMO. Results showed clearly that the gas phase minimum of the cationic $[Al_5O_6H_2Cl_4]^+$ cluster was solvated most effectively having the lowest solvation energy. The structure (**b**) had around 94 kJ mol$^{-1}$ and the structure (**c**) had around 166 kJ mol$^{-1}$ higher solvation energies compared to the structure (**a**).

The minimum energy structure of the cationic $[Al_5O_7H_4Cl_4]^+$ complex consists of two trimeric parts facing in the opposite direction; see structure (**a1**) in Figure 1. The trimeric part on the left-hand side is facing away from the viewer, and the trimeric part on the right-hand side is facing toward the viewer, respectively. The orientation seems to be crucial, since the conformation where both trimeric parts are facing in the same direction is over 40 kJ mol$^{-1}$ higher in energy (PBE-D/TZVPP). The structure of the complex consists of three equivalent three coordinated oxygen atoms and four hydroxo bridges. The aluminum−oxygen bond lengths varied from 1.81 to 1.92 Å, being the longest between center five coordinated aluminum atom and top trivalent oxygen atom, which is linking the two trimeric parts together. The $R_{Al-OH}$ bond lengths varied from 1.86 to 1.88 Å. The structure consists of four chlorido ligands attached to the corner aluminum atoms. The aluminum chlorine bond distances were around 2.08 Å. All aluminum atoms were four coordinated except the shielded center aluminum atom, which had 5-fold coordination. The core structure of the gas phase minimum was also slightly bent, which can be seen from the 130° angle between the oxygen atom of the hydroxo bridge, center five coordinated aluminum atom, and the oxygen atom of the bridging hydroxo group of the other trimeric part.

As a gas phase minimum, the structure (**a1**) was also selected for the CPMD simulations. The selection was based also on the geometrical characteristics of the cluster because the compact trimeric unit is also one of the main components of the tridecameric Keggin cation, which can be viewed as four trimeric $Al_3O(OH)_6(H_2O)_3$ groups linked together at polyhedral edges around the central $Al(O)_4$ unit.[71−74] Furthermore, Keggin cation is widely considered to be one of the main hydrolysis products of aluminum in acidic aquatic solutions.[13−17,28] Thus, the investigation of the stability of the structure (**a1**) in aquatic environments is justified.

The structure (**b1**) differs from the gas phase minimum by the orientation of the trimeric units. In (**b1**), they are both facing toward the viewer, see Figure 1. Furthermore, the two four-rings in both sides of the center five coordinated aluminum atom are in the same horizontal plane. The structure consists of four hydroxo bridges and three trivalent oxygen bridges. The $R_{Al-O}$ and $R_{Al-Cl}$ bond lengths were very similar compared to the equivalent bond distances in the structure (**a1**). The $R_{Al-O}$ bond lengths varied from 1.81 to 1.93 Å. The core of the structure (**c1**) is identical to the structure (**a**), although an additional aqua ligand is attached to the center aluminum atom increasing the coordination of aluminum from four to five. The $R_{Al-OH2}$ bond length was 1.90 Å. In the structure (**d1**), the aqua ligand is attached to the apical position of the left aluminum atom at a distance of 1.93 Å. In this position, the aqua ligand changed also the orientation of the chlorido ligand attached to the same aluminum atom. This can be detected by comparing the bond angles between chloride, corner aluminum, and center trivalent oxygen atoms. The normal angle is around 123.0°; however, due to the aqua ligand, this angle decreased to 107.6°, changing also the bonding of the aluminum atom from tetrahedral to trigonal-bipyramidal.

The solvation of the gas phase optimized structures with COSMO changed the energy differences of the reported structural isomers. According to the results, the two most effectively solvated structures were (**c1**) and (**d1**), as seen in Figure 1. Cavity model results indicated also that the additional water molecule is most likely attached as aqua ligand to the primary hydration shell of structure (**a**) without changing the core. Furthermore, COSMO calculations revealed that the hydration reaction from the structure (**a**) to structure (**d1**) is strongly exothermic. The Gibbs free energy of hydration ($\Delta G_{hyd}$) was under −40 kJ $mol^{-1}$. This shows unambiguously that the structure (**a**) is most likely spontaneously hydrated in aqueous environments.

The search of the ground state structure of the oligomeric $[Al_5O_8H_6Cl_4]^+$ complex was demanding. Not only the amount of structural isomers in the energy surface increased compared to the previous clusters but also three different pentameric aluminum conformations within 2 kJ $mol^{-1}$ (PBE-D/TZVPP) to the ground state structure were detected; see structures (**a2**), (**b2**), and (**c2**) in Figure 2. The common structural characteristic of these conformations was that they all were consisted of four- and six-rings. In addition, the coordination of aluminum varied from four to five.

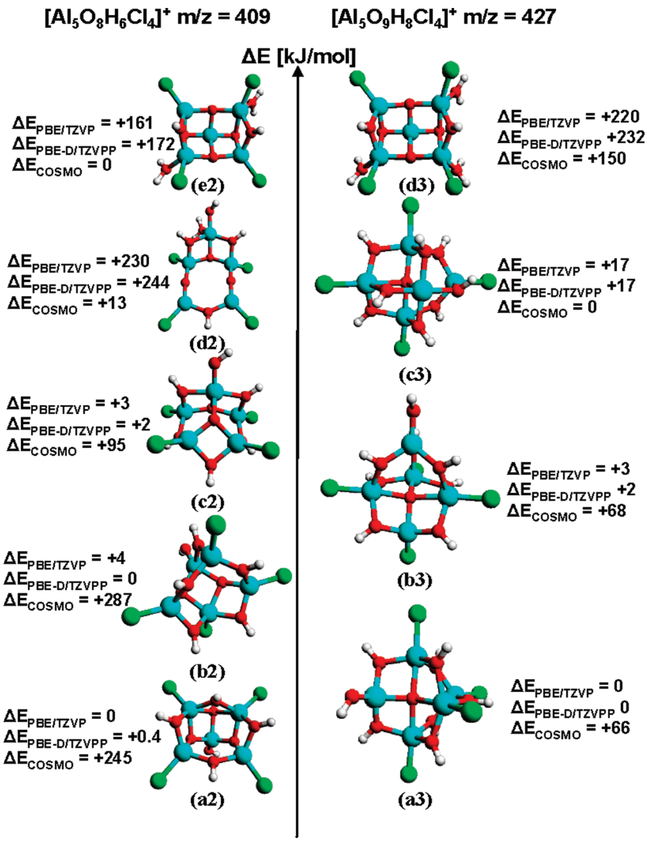


***Figure 2.*** Optimized structures for the cationic $[Al_5O_8H_6Cl_4]^+$ and $[Al_5O_9H_8Cl_4]^+$ aluminum complexes.

The core of the (**a2**) structure consists of a chain of four-rings linked together to a hexagonal shaped ring from both ends with a single hydroxo bridge; see Figure 2. From three to five aluminum atoms were four coordinated, the rest having 5-fold coordination. The two apical aluminum atoms were also joined together with a hydroxo bridge. Structurally, the most interesting detail is the protonated oxygen atom between the two apical aluminum atoms, located behind the hydroxo bridge. It is joining three different aluminum atoms together while being protonated. The $R_{Al-O}$ bond distances varied from 1.68 to 2.14 Å, being the shortest between the center aluminum and oxygen atom of hydroxo ligand. The longest aluminum−oxygen distances were between four coordinated protonated oxygen atoms and the two apical aluminum atoms. The rest of the aluminum−oxygen bond distances varied from 1.79 to 1.94 Å, and $R_{Al-Cl}$ bond lengths varied from 2.08 to 2.09 Å.

The structure (**b2**) can be considered as a combination of tetrameric $[Al_4O_2(OH)_3Cl_3]^{2+}$ and monomeric $[Al(OH)_3Cl]^-$ linked together through hydroxo bridges. The tetrameric part consists of three four-rings of Al−O or Al−OH sides. The monomeric part is then bridging three corner aluminum atoms of the tetrameric part. The core $R_{Al-O}$ bond lengths varied from 1.77 to 1.95 Å. The $R_{Al-OHligand}$ bond distance was around 1.68 Å, whereas $R_{Al-Cl}$ bond lengths varied from 2.08 to 2.10 Å.

Structure (**c2**) consists of two six-rings sharing one Al−O side and linked together with one hydroxo and one trivalent oxygen bridge; see Figure 2. It consists of four equivalent aluminum atoms with 4-fold coordination and one aluminum atom with 5-fold coordination. The core $R_{Al-O}$ bond lengths varied from 1.76 to 1.98 Å whereas the $R_{Al-Cl}$ bond distances were around 2.8 Å. The $R_{Al-OHligand}$ bond distance was around 1.71 Å. Compared to the other two structures close in energy, (**c2**) was most open. Although the structure (**b2**) was slightly lower in energy compared to the other two structures, one cannot say which one of these conformations is the real gas phase minimum. The energy difference of these structures was investigated also with second-order Møller−Plesset perturbation theory with triple $\zeta$ valence with double polarization basis set (TZVPP).[36,42−44] However, all three conformational isomers remained within 5 kJ mol$^{-1}$.

The core structure (**d2**) consists of a trimeric part $[Al_3O(OH)_3(H_2O)Cl_2]^{2+}$ linked together with a dimeric part $[Al_2O_2(OH)Cl_2]^-$ via divalent oxygen bridges. The structure consists of two four-rings and one eight-ring. The $R_{Al-O}$ bond lengths varied from 1.65 to 1.95 Å, whereas $R_{Al-Cl}$ distances varied from 2.07 to 2.13 Å. The shortest (1.65 Å) aluminum−oxygen bond lengths were between aluminum atoms of the dimeric part and divalent oxygen bridges. The core of (**d2**) resembles closely to the core of (**c2**) but is even more open. The core of (**e2**) is almost identical with the core of (**a**). The only difference is the orientation of the two chloride ligands caused by the addition of two aqua ligands to the corner aluminum atoms in the opposite sides of the structure. Due to these ligands, the bond angles between chloride, corner aluminum, and center trivalent oxygen atoms decreased to 108°. In the case of four coordinated aluminum atoms, the equivalent bond angle stayed in 121°. The $R_{Al-OH2}$ bond lengths were around 1.93 Å.

The solvation of the gas phase optimized structures showed that the structure (**c2**) was lower in energy in aquatic solutions compared to the more compact structures (**a2**) and (**b2**); see Figure 2. In addition, the structure (**e2**) was detected to be the liquid phase minimum, although the (**d2**) seemed to be the best solvated structure overall. This is clearly due to the more exposed core structure of (**d2**). According to the COSMO calculations, the hydration reaction from (**d1**) to (**e2**) was exothermic ($\Delta G_{hyd} = -40$ kJ mol$^{-1}$). This indicated that the structure (**a**) experiences most likely at least two spontaneous hydration reactions in aquatic environments. The Gibbs free energy of the hydration of (**d1**) to (**d2**) was also negative ($\Delta G_{hyd} = -27$ kJ mol$^{-1}$), indicating that the core of the structure (**a**) can also open up in aquatic solutions.

The structural isomers of oligomeric $[Al_5O_9H_8Cl_4]^+$, especially the structure (**a3**), (**b3**), and (**c3**) had almost identical adamantane-like cores. The differences in energy were due to the different ligand orientations. The structures (**a3**) and (**c3**) were basically combinations of tetrameric $[Al_4O(OH)_5Cl_3]^{2+}$ and monomeric $[Al(OH)_3Cl]^-$ aluminum species. Their only difference was the additional hydroxo bridge between center aluminum atoms in the tetrameric part of the structure (**c3**). The structure (**a3**) was lacking this bridging agent, and the hydroxo group was attached as a ligand to the center aluminum atom in the left side of the tetrameric part; see Figure 2. This change in the orientation affects to the total energy of the cluster over 15 kJ mol$^{-1}$. The structure of the tetrameric part closely resembled the highly compact cyclic structures of the largest stable configurations of cationic $[Al_4O(OH)_5Cl_4(H_2O)_{0-2}]^+$.[41] The $R_{Al-O}$ bond lengths varied from 1.81 to 2.15 Å, whereas $R_{Al-Cl}$ bond distances varied from 2.09 to 2.13 Å.

The core of the structure (**b3**) can be considered as a combination of the tetrameric $[Al_4O(OH)_4Cl_4]^{2+}$ and monomeric $[Al(OH_4)]^-$ units. The only difference between the structures (**a3**) and (**b3**) was the location of the hydroxo ligand. In structure (**b3**), the hydroxo ligand is attached to the aluminum atom of the monomeric part, whereas in (**a3**) it is attached to the center aluminum in the left side of the tetrameric part, as seen in Figure 2. This, however, affects very mildly the relative energy differences of the structural isomers. The $R_{Al-O}$ bond distances of the (**b3**) varied from 1.81 to 2.06 Å, and $R_{Al-Cl}$ bond lengths varied from 2.08 to 2.14 Å. Every one of the aforementioned structures ((**a3**), (**b3**), and (**c3**)) consists of three aluminum atoms with 5-fold coordination and two aluminum atoms with 4-fold coordination.

The structure (**d3**) had the same core as the structure (**a**) having three aqua ligands attached to the corner aluminum atoms. The bond angles between chloride, corner aluminum, and center trivalent oxygen atoms were approximately 108° for aluminum atoms with 5-fold coordination and 123° for aluminum atoms with 4-fold coordination. The $R_{Al-OH2}$ bond lengths varied from 1.94 to 1.95 Å. Note that all aqua ligands were oriented to the back of the structure.

COSMO calculations changed the ground state conformation from the gas phase minimum (**a3**) to (**c3**). In total, the structure (**c3**) had over 80 kJ mol$^{-1}$ lower solvation energy compared to the gas phase minimum. The Gibbs free energy of hydration from (**e2**) to (**c3**) was strongly exothermic ($\Delta G_{hyd} < -200$ kJ mol$^{-1}$). On the grounds of these COSMO calculations, it is clear that the low coordinated pentameric aluminum clusters are most probably spontaneously hydrated in aquatic solutions. Furthermore, there is a high probability that the core structure of the pentameric complexes changes from compact to more open in liquid conditions. Thus, one of the main goals of the proceeding CPMD part is to investigate the hydrolysis and stability of the aforementioned oligomeric aluminum complexes in aquatic environments.

### 3.2. Car−Parrinello Molecular Dynamics Studies.

In this part, we will concentrate on both the stability of the chosen pentameric aluminum complexes and their hydrolysis reactions in aquatic environments. We note that none of the previous computational studies have focused on the stability and solvation of this kind of oligomeric aluminum compounds. Thus, on the grounds of the results, we are able to improve the prevailing conception of the role of pentameric aluminum clusters in the hydrolysis of aluminum species.

#### 3.2.1. $[Al_5O_6H_2Cl_4]^+$ (a) without vdW Corrections.

The initial system contained 141 explicit water molecules around the cluster (**a**) in a cubic box with 17 Å sides. The total duration of the simulation was 40 ps (ps). During this time, we detected significant changes in the primary hydration shell of the cluster. In addition, due to these spontaneous reactions, the core structure of the pentameric aluminum cluster

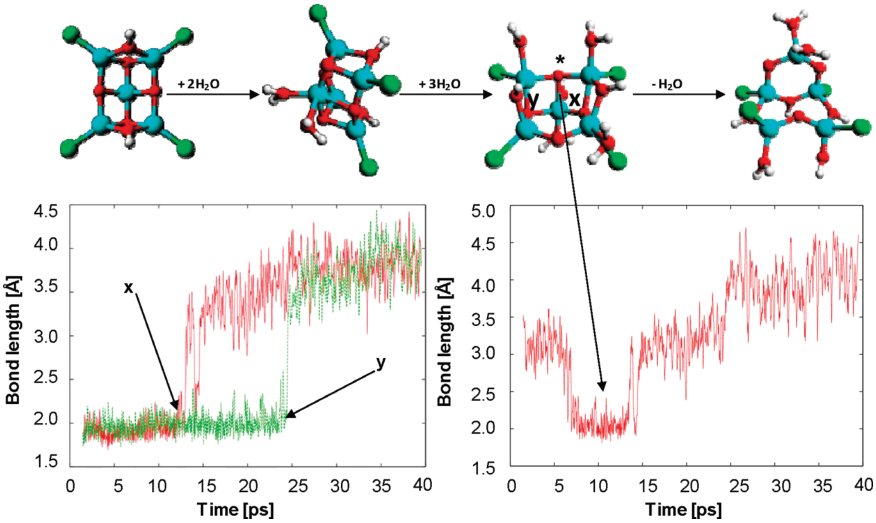


***Figure 3.*** Detected spontaneous associative and dissociative hydration reactions in the simulation of pentameric aluminum complex (**a**). The oscillation of the Al−O bond distances (**x** and **y**) (lower left) and the oscillation of the $Al_{center}$−O* bond (lower right) indicating the metastability of the hexagonal prism-like structure.

changed from a compact cyclic form to an open structure; see Figure 3.

The first two aqua ligands were attached to the center aluminum atom of the ground state structure (**a**). Due to these associative hydration reactions, two of the intramolecular bonds between center aluminum and trivalent oxygen atoms on the left and right side of the core of cluster (**a**) were broken; see Figure 3. Bond dissociation was then followed by three associative hydration reactions to the corner aluminum atoms of the structure. We note that all of these hydration reactions occurred during the first 7 ps. During these reactions, the coordination of aluminum increased from four to five. However, one of the corner aluminum atoms was not hydrated, maintaining a 4-fold coordination. The lifetime of this highly symmetric and coordinated intermediate structure was from 7 to 8 ps; see the valley in the lower right corner in Figure 3. In addition, the core of this intermediate structure closely resembled the core of (**a2**). Before the addition reaction of a new aqua ligand to the open coordination position of the four coordinated aluminum atom, the dissociation of one of the already attached aqua ligands occurred (12 ps), leading to a opening of the structure.

The mechanism for the structural reorganization was as follows; first, the Al−O bonds between the center aluminum and trivalent oxygen atom (between bonds **x** and **y**) and **x** were broken, following the breakage of the bond **y** around 23 ps; see Figure 3. The newly formed structure consisted of one six- and one eight-rings, as seen from the final structure in Figure 3. During these reactions, the coordination of aluminum decreased back to four. The final open structure stayed intact for the last 15 ps without any hydration reactions. The opening of the structure indicated clearly that the gas phase ground state structure was unstable in aquatic solutions, which agrees well with the earlier COSMO findings.

*3.2.2. $[Al_5O_6H_2Cl_4]^+$ (a) with Empirical vdW Corrections.* The effect of the van der Waals corrections was investigated by employing empirical parameters to describe the vdW interactions within DFT-PBE. The starting geometry of the system was identical compared to the previous simulation. The total duration of the simulation was around 30 ps. During the simulation, six spontaneous associative hydration reactions occurred. The mechanism for the reactions was as follows; the first two water molecules were attached to the center aluminum atom raising its coordination temporarily from four to six; second, two of the corner aluminum atoms were successively hydrated raising the coordination of aluminum from four to five; see Figure 4. The fifth additional water molecule was attached to the same corner aluminum atom as the fourth aqua ligand (circulated in the middle structure in Figure 4.) raising its coordination temporarily from five to six. Finally, the sixth aqua ligand was attached to the same aluminum atom as the third aqua ligand leading to a 6-fold coordination of aluminum. During these reactions, the structure experienced also two intramolecular aluminum−oxygen bond breakings leading to an opening of the structure.

The mechanism of the spontaneous hydration reactions occurred with similar mechanism compared to the same simulation without empirical vdW corrections. The first four hydration reactions followed the same path with only one exception: the fourth aqua ligand (circulated in the third structure in Figure 4) was attached to the equatorial position herein, whereas it was attached to the axial position in the previous simulation. In addition, the first two intramolecular Al−O bond breakings followed an identical path compared to the previous simulation. The most significant difference between the simulations mentioned above was that the third intramolecular Al−O bond **y** was broken in the simulation without empirical corrections whereas it stayed intact during the vdW-corrected simulation, preventing the structure to fully open. During this simulation, spontaneous dissociation of one of the chlorido ligands was also observed. After the breakage of the aluminum-chloride bond, the newly formed chloride ion diffused and solvated around 5 Å off the cluster, decreasing the coordination of aluminum to five. The detected dissociation reaction is consistent with the Al−Cl bond dissociation energy barriers determined in our previous computational studies.[67]

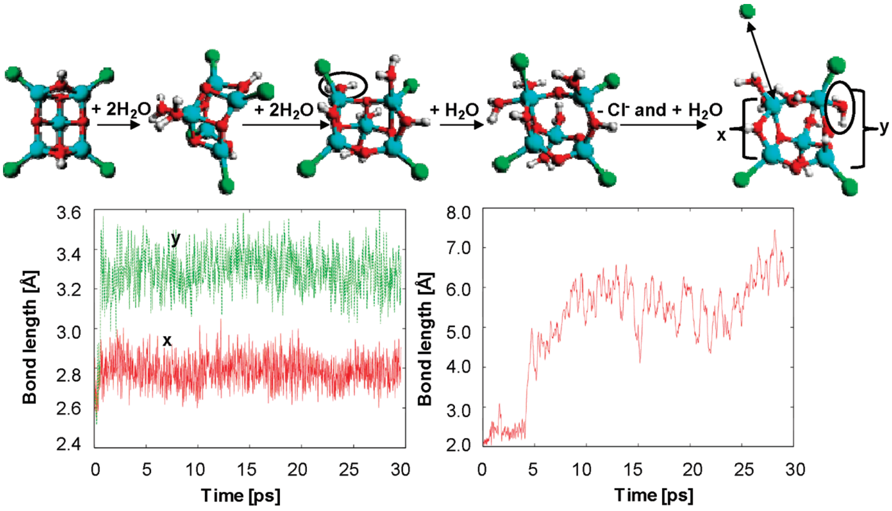


***Figure 4.*** Detected spontaneous associative hydration reactions and the dissociation of the chloride ion. The oscillation of the Al−Al distances (lower left) of the corner aluminum atoms in the eight-ring of the cage-like structure in the vdW-corrected simulation indicating the stability of the cluster and the Al−Cl bond oscillation and the bond breaking in the lower right corner.

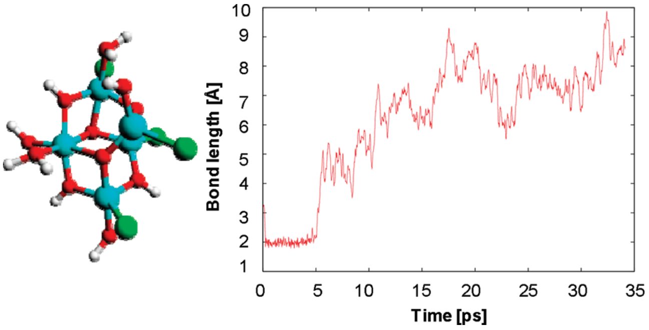


***Figure 5.*** Final geometry in the simulation of the pentameric complex (**b**) (left) and the oscillation of the Al−O bond distance pointing the dissociation of one of the aqua ligands from the primary hydration shell (right).

We note that the first five hydration reactions and the dissociation of the chlorido ligand occurred during the first 5 ps of the simulation, indicating faster reaction speeds compared to the previous simulation. The hydration seemed to proceed further even with shorter simulation times. This agrees well with the results of the liquid water investigation of Lin et al.[75] They postulated that the vdW-corrected DFT-BLYP produces improved structural and dynamical properties of the liquid water increasing the self-diffusion coefficient and making the water softer and more liquid-like.[75] The final sixth additional aqua ligand in the equatorial position (circulated in Figure 4) was attached to the primary hydration shell of the cage-like complex around 28 ps of the simulation.

The final cage-like structure consists of one four-, one six-, and one eight-ring; see Figure 4. It closely resembles the intermediate structure after the first two internal aluminum−oxygen bond breakages in the previous simulation. The oscillation (no visible drift) of the corner Al−Al distances of the eight-ring revealed that, although the structure was asymmetric, it remained intact during the last 25 ps of the simulation. The core remained intact also after the final sixth spontaneous hydration reaction. The structural reorganization and the opening of the cluster strengthened our previous conclusions that the original structure is not stable in aquatic environments. The final sum molecular formulas of the simulations of the ground state structure (**a**) were $[Al_5O_4(OH)_3(H_2O)_5Cl_3]^+$ with and $[Al_5O_3(OH)_4(H_2O)_3Cl_4]^+$ without vdW corrections.

### 3.2.3. $[Al_5O_6H_2Cl_4]^+$ (b) without vdW Corrections.

The structure (**b**) was also surrounded by 141 water molecules to a cubic cell of 17 Å sides. The total duration of the simulation was around 33 ps, in which the core of the gas phase optimized structure experienced six spontaneous hydration reactions. We note that all of these associative hydration reactions occurred in the first 4 ps of the simulation. In addition, one of the aqua ligands dissociated from the primary hydration shell back to the solution around 5 ps; see Figure 5. The sum reaction mechanism for the hydration reactions was as follows

$$[Al_5O_4(OH)_2Cl_4]^+ + 5H_2O \rightarrow [Al_5O_4(OH)_2(H_2O)Cl_4]^+ + 4H_2O \rightarrow [Al_5O_4(OH)_2(H_2O)_2Cl_4]^+ + 3H_2O \rightarrow [Al_5O_4(OH)_2(H_2O)_3Cl_4]^+ + 2H_2O \rightarrow [Al_5O_4(OH)_2(H_2O)_4Cl_4]^+ + H_2O \rightarrow [Al_5O_3(OH)_4(H_2O)_4Cl_4]^+ \quad (3)$$

The first two aqua ligands were attached to the aluminum atom without chlorido ligand, increasing the coordination of aluminum from four to six; see structure (**b**) in Figure 1. This was followed by the hydration of the two top corner aluminum atoms of the cubane-like moiety. During these reactions, the coordination of the corner aluminum atoms raised from four to five. After the fourth spontaneous associative hydration reaction, the intramolecular Al−O bond between the aforementioned six coordinated aluminum and trivalent oxygen atom was broken, slightly opening the structure and decreasing the coordination of aluminum from six to five. This was then followed by the hydration of the remaining four coordinated aluminum atoms. After the associative hydration reactions, the fourth additional aqua

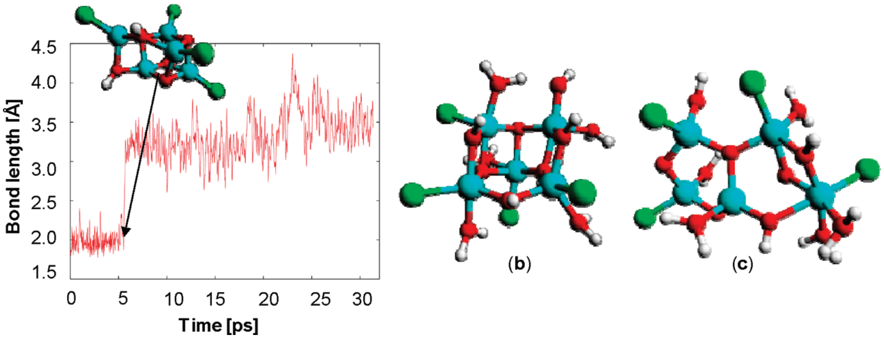


***Figure 6.*** Al−O bond (arrow) oscillation indicating the opening of the structure (**b**) in vdW-corrected simulation (left), the final structure of (**b**) in vdW-corrected simulation (middle), and the final structure of (**c**) in the vdW-corrected simulation (right).

ligand dissociated off from the cluster; see Figure 5. During this simulation, one of the aqua ligands was also deprotonated. The dissociated proton was then captured by one of the core oxygen atoms. This intramolecular proton transfer led to a formation of bridging hydroxyl group and additional hydroxo ligand; see the final formula in eq 3.

The final structure after the last step of the simulation consisted of three equivalent oxygen bridges with 3-fold coordination and three hydroxyl bridges. Furthermore, the coordination of aluminum varied from four to six. Besides the hydration reactions, the structure (**b**) experienced only minor structural changes during the simulation. The sum molecular formula of the final structure was $[Al_5O_3(OH)_4(H_2O)_4Cl_4]^+$, as seen in Figure 5.

*3.2.4. $[Al_5O_6H_2Cl_4]^+$ (b) and (c) with Empirical vdW Corrections.* The simulation was performed as before and the total duration was around 32 ps. During this time, the cluster experienced seven spontaneous hydration reactions. The first six associative hydration reactions occurred during the first 5 ps, and the last seventh aqua ligand was attached to the primary hydration shell of the cluster around 21 ps. The mechanism for the reactions went as follows; first, two aqua ligands were attached to the aluminum atom without chlorido ligand, followed by three successive hydration reactions of the four coordinated aluminum atoms. The sixth additional aqua ligand, however, was attached to the corner aluminum atom of the cubane-like moiety with 5-fold coordination. Finally, the seventh associative hydration reaction occurred in the only remaining four coordinated aluminum atoms; see Figure 6.

During the simulation, the core of the complex (**b**) experienced significant topological changes, triggered by the breaking of the intramolecular Al−O bond (arrow in Figure 6) around 6 ps. The bond was broken between the center aluminum atom without chlorido ligand and the trivalent oxygen bridge facing toward the viewer; see Figure 1. The newly formed structure consisted of a chain of four $Al_2O_2$-rings linked together to a hexagonal shaped ring from both ends of the chain with a single hydroxo bridge. The core of this highly symmetric structure closely resembled the core of the gas phase optimized (**a2**) and was almost identical with the core of the metastable (~8 ps) intermediate structure in the simulation of the structure (**a**) without empirical corrections; see Figures 2 and 4. In the case of the newly formed hexagonal prism-like structure, however, the structure stayed intact during the rest of the 25 ps of the simulation, indicating the stability of the structure in aquatic environments; see Figure 6. This is due to the different ligand orientation compared to the metastable structure in the simulation of the structure (**a**).

The most noticeable differences were the following: first, the center aluminum atom in the structure is six coordinated whereas in the metastable structure it was five coordinated; second, one of the chlorido ligands is attached to the center aluminum atom while chlorido ligands were attached only to the corner aluminum atoms in the intermediate structure. The differences in the stability can then be explained by decreased ligand repulsion and structural straining in the complex. The final structure ((**b**) in Figure 6) closely resembles also the hexagonal prism-like crystal structure of Harlan et al.[76] However, their crystal structure analogue contained six aluminum atoms instead of five.[76] The final sum molecular formula of the newly formed pentameric aluminum complex can be written as $[Al_5O_3(OH)_4(H_2O)_6Cl_4]^+$, indicating further hydrolysis compared to the identical simulation without empirical vdW corrections.

The stability of the structure (**c**) was investigated only in a simulation of 141 explicit water molecules with empirical van der Waals corrections. The structure experienced six spontaneous associative hydration reactions during the 41 ps production run. The only topological change of the cluster during the simulation besides aforementioned hydration reactions was the opening of the Al−O bond between center aluminum atom and the trivalent oxygen bridge of the trimeric moiety. This was caused by intramolecular proton diffusion from one of the aqua ligands in the primary hydration shell to the trivalent oxygen atom. Otherwise, the core of the structure remained intact during the simulation, indicating that the structure is rather stable in liquid conditions; see Figure 6. The final sum molecular formula of the cluster can be written as $[Al_5O_3(OH)_4(H_2O)_5Cl_4]^+$.

*3.2.5. $[Al_5O_7H_4Cl_4]^+$ (a1) without vdW Corrections.* The initial system contained 169 explicit water molecules around the cluster (**a1**) in a cubic box with 18 Å sides. The total duration of the simulation was around 44 ps. During this time, the topology of the cluster changed significantly from highly symmetric cyclic to an open structure. A pentameric aluminum complex (**a1**) experienced in a total of four associative hydration reactions. In addition, the breaking of the intramolecular Al−O bond between the center aluminum atom and the topmost trivalent bridging oxygen atom occurred; see Figure 1. The final structure of the simulation

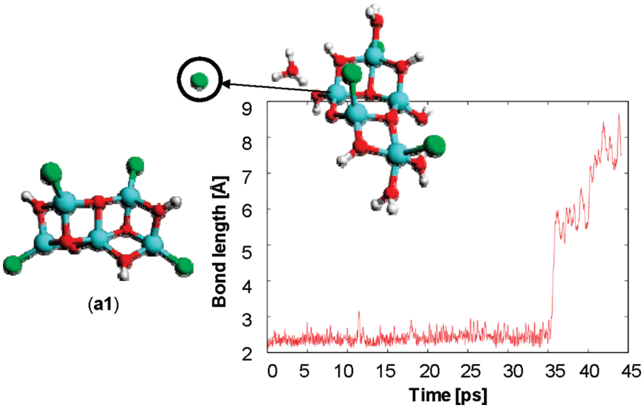


***Figure 7.*** Initial gas phase structure (left), the Al–Cl bond oscillation indicating the dissociation of the chlorido ligand, and the final geometry of the simulation (right).

can be considered as a combination of adamantane-like tetrameric moiety $[Al_4O_3(OH)_3(H_2O)_2Cl_2]^+$ and a monomeric moiety $[Al(OH)(H_2O)_2Cl]^+$; see Figure 7. We note that the four hydration reactions and the intramolecular bond breakage occurred during the first 9 ps.

During the simulation (~35 ps), the dissociation of one of the chlorido ligands was also observed; see Figure 7. After the breakage of the aluminum–chlorine bond, the newly formed chloride ion (circulated) was diffused further and solvated 6 to 8 Å off the cluster, decreasing the coordination of aluminum from five to four. The breaking of one of the intramolecular Al–O bonds between the center aluminum atom and oxygen atom of the bridging hydroxo group in the left trimeric moiety was also detected, leading to a formation of an additional hydroxo ligand. Akitt et al. suggested that this kind of hydroxo ligand formation can trigger the polymerization of monomeric aluminum species to dimeric complexes.[3] In the case of monomeric species, the formation of the hydroxo ligand is due to proton transfer, but in the case of oligomeric aluminum complexes, such as (**a1**), the formation can be caused also by intramolecular topological changes. Detected associative hydration reactions and changes in the core structure indicated clearly that the original structure was unstable in aquatic environments. In addition, the formation of hydroxo ligand strengthened the conclusion, indicating that the structure is a likely candidate for further polymerization reactions. The final sum molecular formula of the cluster can then be written as $[Al_5O_3(OH)_4(H_2O)_4Cl_3]^{2+}$. The charge was due to the dissociation of the chlorido ligand.

*3.2.6. $[Al_5O_7H_4Cl_4]^+$ (a1) with Empirical vdW Corrections.* The stability of (**a1**) was also tested in an identical vdW-corrected simulation of 32 ps. During this time, the gas phase optimized structure (**a1**) experienced four hydration reactions and three intramolecular bond breakings, leading to a significant topological changes and the opening of the ground state structure. The intramolecular Al–O bond breakings followed the same mechanism compared to the aforementioned simulation without empirical vdW corrections. The final structure of the simulation consisted of two Al–O four-rings linked together from both ends by an eight-ring, as seen in Figure 8.

Transformation of one of the bridging hydroxo groups to a hydroxo ligand was also detected, as in the previous simulation without vdW corrections. The breaking of the Al–O bond, which triggered the conversion, occurred around 16 ps of the simulation. This strengthened the previous conclusions that the structure (**a1**) is unstable in aquatic environments. After the structural reorganization, the newly formed open complex stayed intact experiencing only one attempt for the intramolecular Al–O bond dissociation between 16 and 23 ps of the simulation; see Figure 8. The final sum molecular formula of the cluster can be written as $[Al_5O_3(OH)_5(H_2O)_3Cl_4]^0$. The neutral charge of the complex is due to the proton transfer reactions between the cluster and surrounding water molecules.

During these seven individual simulations, we detected several (4/7) structural rearrangements of the compact symmetric structures to an open or cage-like structure. The open structure and highly symmetric hexagonal prism-like structures were found to be dominating geometries for the pentameric aluminum complexes in liquid conditions. Furthermore, several spontaneous associative hydration reactions occurred in every simulation, indicating that the structures detected in the ESI MS experiments are low coordinated.[13–16] The same phenomenon was also detected in our previous CPMD studies of the dimeric aluminum chlorohydrates in aquatic environments.[66] An interesting observation of these simulations was that the hydrolysis proceeded further in the simulations with empirical long-range vdW corrections compared to the simulations without corrections. In addition, the hydration reactions occurred faster, enabling shorter simulation times in the vdW-corrected simulations. We note that the majority of the Al–O bond lengths detected during the CPMD part of this study were within the typical Al–O single bond distance range.[67,68]

*3.2.7. Solvation of the Clusters.* In this section, we concentrate on the structural characteristics of the surrounding water. This was done by investigating the total HO– and OO– radial distribution functions from the cationic $[Al_5O_7H_4Cl_4]^+$ ($m/z = 391$) system with and without empirical van der Waals corrections. The first peak in HO–RDF with the distance of 0.87–1.16 Å, maximum at 0.99 Å, corresponds to the OH distances in water molecules and in the cluster, Figure 9. The second peak at the distance of 1.29–2.40 Å gives us the total amount of acceptor and donor hydrogen bonds in the systems. The maximum of the second peak of around 1.78 Å for both systems is close to the experimental value (1.8 Å) of the hydrogen bond.[77,78] The shapes and positions of the g(O,H) and g(O,O) RDFs in both systems are in good agreement with the water diffraction data of Soper et al.[77,78] In OO–RDF, the first peak at the distance of 2.33–3.31, maximum at 2.73 Å, and the second peak at the distance 3.51–5.51, maximum at 4.49 Å, are almost identical to the experimental values.[77,78] The results are also in very good agreement with the CPMD results of Sillanpää et al.[21] and with the results of liquid water studies of Kuo et al.[79]

It is known that the PBE density functional tends to slightly overstructure pair correlation functions and, in addition, the self-diffusion coefficient is much smaller than in experiments

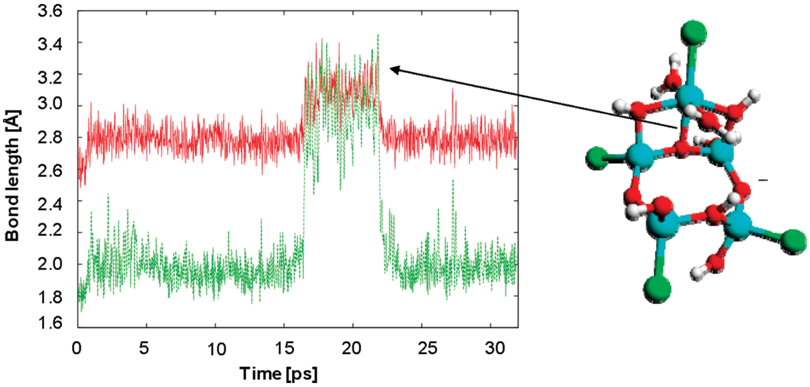


***Figure 8.*** Oscillations of the Al−O (green) bond between the topmost aluminum atom and the center trivalent oxygen atom and Al−Al (red) distance between the topmost aluminum and the corner aluminum atom in the same four-ring are indicating the stability of the final open structure (right). The peak in the figure corresponds to the intramolecular Al−O bond breaking.

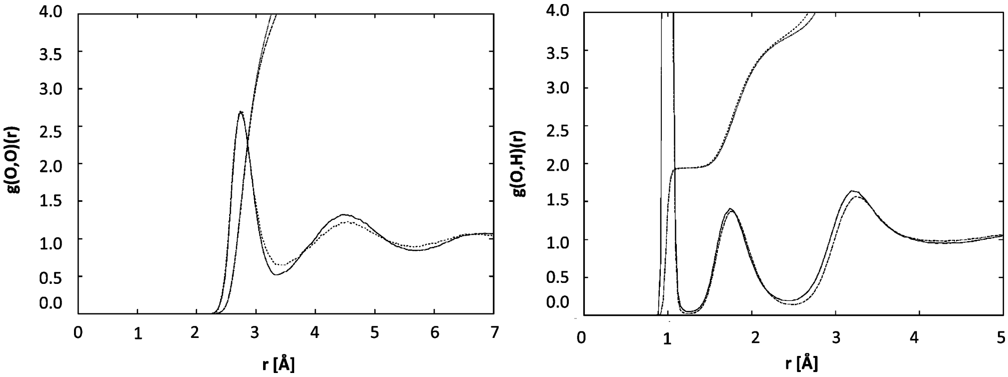


***Figure 9.*** Differences in the total OH− and OO−RDFs of the $[Al_5O_7H_4Cl_4]^+$ ($m/z = 391$) simulation with and without empirical vdW corrections. Dashed line indicates the system with empirical corrections in g(O,O)(r) and in g(O,H)(r); the solid line indicates the vdW-corrected system. The upper integral belongs to the vdW-corrected system, and the lower belongs to the system without empirical corrections.

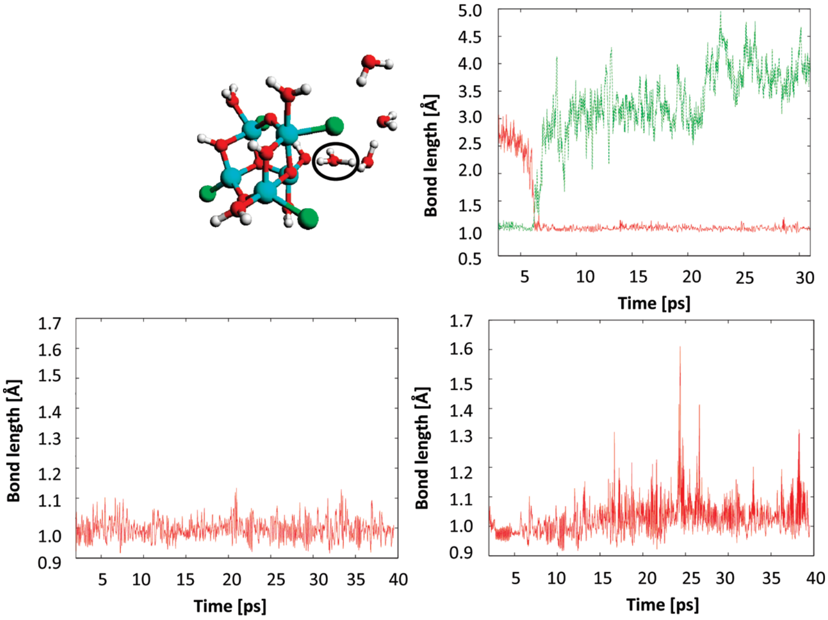


***Figure 10.*** Oscillation of the covalent O−H bond of the aqua ligand (lower right) and the solvent water molecule (lower left). The formation of the hydronium ion (circulated) to the secondary hydration shell is in the upper left corner and the oscillation of the O−H distances indicating the intramolecular proton transfers from aqua ligand (green) to the trivalent unprotonated core oxygen atom (red) is in upper right corner.

making the water sluggish.[80,81] However, according to the findings of Lin et al., the usage of empirical van der Waals corrections in CPMD decreases the difference in the self-diffusion coefficient between computations and experiments.[75] Although the diffusion coefficient was not measured in this study, the g(O,O) of the vdW-corrected system displayed softer structure (dashed line in Figure 9) for the water compared to the DFT-PBE simulation and closer to the experimental results of Soper et al.[77,78] Slow diffusion should not be a problem in the systems without empirical corrections either due to the rather long simulation times and higher temperatures (350 K).[66]

During the simulations, we detected several attempts of the protons of aqua ligands to jump to the surrounding water, indicating the acidity of the pentameric aluminum complexes. In the majority of the cases, the protons stayed intact in the cluster without diffusion to the surrounding solvent. In almost every simulation, however, transient proton jumps were detected. The mechanism for these proton transfers was always the same: a nearby solvent molecule captured the proton from the aqua ligand or bridging hydroxo group; see Figure 10. As a result, the hydronium ion ($[H_3O]^+$) was formed in the secondary hydration shell. This is in good agreement with the studies of Tuckerman et al.[82,83] The acidity of the aqua ligands was also seen by observing the covalent oxygen−hydrogen bond distances. For the solvent water molecule, this distance oscillated from 0.92 to 1.1 Å with an average around 1.0 Å; whereas for the aqua ligands, the distance varied from 0.92 to 1.6 Å; see Figure 10. During these simulations, we detected also several (4/7 cases) intramolecular proton transfers from aqua ligands to the trivalent unprotonated core oxygen atoms.

## 4. Conclusions

We used a static quantum chemical and Car−Parrinello molecular dynamics (CPMD) approach to investigate the structural characteristics, the stability, and the hydrolysis of pentameric aluminum complexes in both gas phase and in aquatic environments. We tested the accuracy of several generalized gradient approximation (GGA) and hybrid exchange-correlation functionals in reference to the second-order Møller−Plesset perturbation theory (MP2). The PBE density functional with empirical van der Waals corrections gave the most coherent results and was selected (PBE-D/TZVPP) for the gas phase conformational analysis. In total, we analyzed hundreds of experimentally detected (ESI MS) structural isomers to find the ground state structures for cationic $[Al_5O_6H_2Cl_4]^+$, $[Al_5O_7H_4Cl_4]^+$, $[Al_5O_8H_6Cl_4]^+$, and $[Al_5O_9H_8Cl_4]^+$ complexes. The analysis revealed that the minimum energy structure was always changed when switching from one cluster to another, indicating that there is not any structural explanation for the crystallization detected in ESI MS studies of Sarpola et al.[13−16,32]

Conductor-like screening model (COSMO) was used to investigate the stability of the gas phase optimized structures in aquatic environments. COSMO calculations revealed that the low coordinated pentameric aluminum clusters are spontaneously hydrated in aquatic solutions. Furthermore, COSMO results indicated that there is a high possibility for the gas phase ground state structures to open up in aquatic solutions, especially when the amount of aqua ligands increases in primary hydration shell. The hydration reactions from $[Al_5O_6H_2Cl_4]^+$ to $[Al_5O_9H_8Cl_4]^+$ were also detected to be strongly exothermic ($\Delta G < 0$).

We performed seven different CPMD simulations to reveal the stability of the chosen pentameric clusters in aquatic solutions. The effect of the long-range empirical van der Waals corrections (-D) was also tested, employing two identical simulations one with and one without the corrections. During these simulations, we detected several spontaneous associative hydration reactions in the primary hydration shell, indicating that the structures detected in the ESI MS experiments are low coordinated. Hydrolysis was also detected to proceed further in vdW-corrected simulations. In addition, the hydration reactions occurred faster, enabling shorter simulation times in the DFT-PBE-D simulations.

During most of the simulations, the chosen structures experienced significant topological changes. In four out of seven cases, the compact cyclic structure was opened leading to a formation of an open or cage-like structure. The open structure, cage-like, and highly symmetric hexagonal prism-like structures were found to be dominating geometries for the pentameric aluminum complexes in liquid environments. Several spontaneous associative hydration reactions were also detected in every simulation. Although we did not find any unique structure for the clusters, the dynamics and the ability of the pentameric complexes to transform in aquatic environments was seen within the picosecond time scales of the simulations. In addition, the structural reorganization and the composition of the open structures (e.g., (**a1**) simulations) indicated that the $Al_5$-clusters are most likely only metastable intermediate forms in the aluminum salt hydrolysis during coagulation, which is in good agreement with the ESI MS findings of Zhao et al.[17]

**Acknowledgment.** The authors thank the graduate school, LASKEMO (Academy of Finland), and the Tauno Tönning foundation for financial support and the IT Center for Science (CSC) for computational resources and expert service. The authors also thank Professor Jürg Hutter for providing a new version of the CPMD code and Dr. Atte Sillanpää for the guidance with the code and useful discussions. Dr. Arja Sarpola is also acknowledged for general support concerning ESI MS results.

## References

(1) Brosset, C. *Acta Chem. Scand.* **1952**, *6*, 910.

(2) Van Cauwelaert, F. H.; Bosman, H. *J. Rev. Chim. Minér.* **1969**, *6* (3), 611.

(3) Akitt, J. W.; Greenwood, N. N.; Khandelwal, B. L.; Lester, G. D. *J. Chem. Soc., Dalton Trans.: Inorg. Chem.* **1972**, *5*, 604.

(4) Bottero, J. Y.; Cases, J. M.; Fiessinger, F.; Poirier, J. E. *J. Phys. Chem.* **1980**, *84* (22), 2933.

(5) Akitt, J. W.; Elders, J. M. *J. Chem. Soc., Dalton Trans.* **1988**, *5*, 1347.

(6) Fu, G.; Nazar, L. F.; Bain, A. D. *Chem. Mater.* **1991**, *3*, 602.

(7) Nazar, L. F.; Fu, G.; Bain, A. D. *J. Chem. Soc., Chem. Commun.* **1992**, *1992*, 251.

(8) Allouche, L.; Gérardin, C.; Loiseau, T.; Férey, G.; Taulelle, F. *Angew. Chem., Int. Ed.* **2000**, *39* (3), 511.

(9) Allouche, L.; Taulelle, F. *Inorg. Chem. Commun.* **2003**, *6*, 1167.

(10) Shafran, K. L.; Perry, C. C. *Dalton. Trans.* **2005**, *12*, 2089.

(11) Johansson, G. *Acta Chem. Scand.* **1960**, *14* (3), 771.

(12) Seichter, W.; Mögel, H.-J.; Brand, P.; Salah, D. *Eur. J. Inorg. Chem.* **1998**, *6*, 795.

(13) Sarpola, A.; Hietapelto, V.; Jalonen, J.; Jokela, J.; Laitinen, R. S. *J. Mass Spectrom.* **2004**, *39*, 423.

(14) Sarpola, A.; Hietapelto, V.; Jalonen, J.; Jokela, J.; Laitinen, R. S.; Rämö, J. *J. Mass Spectrom.* **2004**, *39*, 1209.

(15) Sarpola, A. T.; Hietapelto, V. K.; Jalonen, J. E.; Jokela, J.; Rämö, J. H. *Int. J. Environ. Anal. Chem.* **2006**, *86* (13), 1007.

(16) Sarpola, A. T.; Saukkoriipi, J. J.; Hietapelto, V. K.; Jalonen, J. E.; Jokela, J. T.; Joensuu, P. H.; Laasonen, K. E.; Rämö, J. H. *Phys. Chem. Chem. Phys.* **2007**, *9*, 377.

(17) Zhao, Z.; Liu, H.; Qu, J. *J. Colloid Interface Sci.* **2009**, *330* (1),), 105.

(18) Gale, J. D.; Rohl, A. L.; Watling, H. R.; Parkinson, G. M. *J. Phys. Chem. B* **1998**, *102*, 10372.

(19) Martinez, A.; Tenorio, F. J.; Ortiz, J. V. *J. Phys. Chem. A* **2001**, *105*, 8787.

(20) Martinez, A.; Tenorio, F. J.; Ortiz, J. V. *J. Phys. Chem. A* **2001**, *105*, 11291.

(21) Sillanpää, A. J.; Päivärinta, J. T.; Hotokka, M. J.; Rosenholm, J. B.; Laasonen, K. E. *J. Phys. Chem. A* **2001**, *105*, 10111.

(22) Martinez, A.; Sansores, L. E.; Salcedo, R.; Tenorio, F. J.; Ortiz, J. V. *J. Phys. Chem. A* **2002**, *106*, 10630.

(23) Martinez, A.; Tenorio, F. J.; Ortiz, J. V. *J. Phys. Chem. A* **2003**, *107*, 2589.

(24) Bock, C. W.; Markham, G. D.; Katz, A. K.; Glusker, J. P. *Inorg. Chem.* **2003**, *42*, 1530.

(25) Gowtham, S.; Lau, K. C.; Deshpande, M.; Pandey, R.; Gianotto, A. K.; Groenewold, G. S. *J. Phys. Chem. A* **2004**, *108*, 5081.

(26) Ahu Akin, F.; Jarrold, C. C. *J. Chem. Phys.* **2004**, *120* (18), 8698.

(27) Pophristic, V.; Klein, M. L.; Holerca, M. N. *J. Phys. Chem. A* **2004**, *108*, 113.

(28) Pophristic, V.; Balagurusamy, V. S. K.; Klein, M. L. *Phys. Chem. Chem. Phys.* **2004**, *6*, 919.

(29) Ikeda, T.; Hirata, M.; Kimura, T. *J. Chem. Phys.* **2006**, *124*, 074503-1.

(30) Akitt, J. W.; Elders, J. M. *J. Chem. Soc., Dalton Trans.* **1988**, *5*, 1347.

(31) Akitt, J. W.; Elders, J. M.; Fontaine, X. L. R.; Kundu, A. K. *J. Chem. Soc., Dalton Trans.* **1989**, *10*, 1889.

(32) Sarpola, A. The Hydrolysis of Aluminum, A Mass Spectrometric Study. Ph.D. Dissertation, University of Oulu, Oulu, Finland, Acta Univ. Oul., 2007; C 279, pp 1−104.

(33) Das, U.; Raghavachari, K.; Jarrold, C. C. *J. Chem. Phys.* **2005**, *122*, 014313-1.

(34) Das, U.; Raghavachari, K. *J. Chem. Theory Comput.* **2008**, *4*, 2011.

(35) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77* (18), 3865.

(36) Schäfer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100* (8), 5829.

(37) Ahlrichs, R.; Armbruster, M. K.; Bär, M.; Baron, H.-P.; Bauernschmitt, R.; Böcker, S.; Crawford, N.; Deglmann, P.; Ehrig, M.; Eichkorn, K.; Elliot, S.; Furche, F.; Haase, F.; Häser, M.; Hättig, C.; Hellweg, A.; Horn, H.; Huber, C.; Huniar, U.; Kattannek, M.; Köhn, A.; Kölmel, C.; Kollwitz, M.; May, K.; Nava, P.; Ochsenfeld, C.; Öhm, H.; Patzelt, H.; Rappoport, D.; Rubner, O.; Schäfer, U.; Sierka, M.; Treutler, O.; Utterreiner, B.; Arnim, M.; Weigend, F.; Weis, P.; Weiss, H. Turbomole Program Package for ab initio Electronic Structure Calculations, Users Manual, University of Karlsruhe and Forschungszentrum, Karlsruhe, GmbH, 1989−2007, TURBOMOLE GmbH, 2009; pp 1−341.

(38) (a) Eichkorn, K.; Treutler, H.; Öhm, H.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *240*, 283. (b) Eichkorn, K.; Treutler, H.; Öhm, H.; Häser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *242* (6), 652.

(39) Grimme, S. *J. Comput. Chem.* **2004**, *25* (12), 1463.

(40) Grimme, S. *J. Comput. Chem.* **2006**, *27* (15), 1787.

(41) Saukkoriipi, J.; Sillanpää, A.; Laasonen, K. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3785.

(42) Weigend, F.; Häser, M. *Theor. Chem. Acc.* **1997**, *97* (1−4), 331.

(43) Weigend, F.; Häser, M.; Patzelt, H.; Ahlrichs, R. *Chem. Phys. Lett.* **1998**, *294* (1−3), 143.

(44) Weigend, F.; Furche, F.; Ahlrichs, R. *J. Chem. Phys.* **2003**, *119* (24), 12753.

(45) Becke, A. D. *Phys. Rev. A* **1988**, *38* (6), 3098.

(46) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37* (2), 785.

(47) Becke, A. D. *J. Chem. Phys.* **1993**, *98* (7), 5648.

(48) Becke, A. D. *J. Chem. Phys.* **1996**, *104* (3), 1040.

(49) Grimme, S. *J. Chem. Phys.* **2006**, *124*, 034108-1.

(50) Schwabe, T.; Grimme, S. *Phys. Chem. Chem. Phys.* **2007**, *9*, 3397.

(51) Schäfer, A.; Klamt, A.; Sattel, D.; Lohrenz, J. C. W.; Eckert, F. *Phys. Chem. Chem. Phys.* **1993**, *2*, 2187.

(52) Klamt, A.; Schüürmann, G. *J. Chem. Soc., Perkin Trans.* **1993**, *2*, 799.

(53) Burgess, J. Metal Ions in Solution; John Wiley & Sons: Chichester, Sussex, England, 1978; pp 179−186.

(54) Tissandier, M. D.; Cowen, K. A.; Feng, W. Y.; Gundlach, E.; Cohen, M. H.; Earhart, A. D.; Tuttle, T. R., Jr.; Coe, J. V. *J. Phys. Chem. A* **1998**, *102* (40), 7787.

(55) Coskuner, O.; Jarvis, E. A. A. *J. Phys. Chem. A* **2008**, *112* (12), 2628.

(56) Toukan, K.; Rahman, A. *Phys. Rev. B* **1985**, *31*, 2643.

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

(58) Parrinello, M.; Hutter, J.; Marx, D.; Focher, P.; Tuckerman, M.; Andreoni, W.; Curioni, A.; Fois, E.; Roetlisberger, U.; Giannozzi, P.; Deutsch, T.; Alavi, A.; Sebastiani, D.; Laio, A.; VandeVondele, J.; Seitsonen, A.; Billeter, S. CPMD, Car−Parrinello Molecular Dynamics. http://www.cpmd.org/, Copy-

right IBM Corp 1990−2008, Copyright MPI für Festkörperforschung Stuttgart 1997−2001; pp 1−258.

(59) Nose, S. *J. Phys. Chem.* **1984**, *81* (1), 511.

(60) Nose, S. *Mol. Phys.* **1984**, *52* (2), 255.

(61) Hoover, W. G. *Phys. Rev. A* **1985**, *31* (3), 1695.

(62) Martyna, G. J. *Phys. Rev. E* **1994**, *50* (4), 3234.

(63) Vanderbilt, D. *Phys. Rev. B* **1990**, *41* (11), 7892.

(64) Laasonen, K.; Car, R.; Lee, C.; Vanderbilt, D. *Phys. Rev. B* **1991**, *43* (8), 6796.

(65) Laasonen, K.; Pasquarello, A.; Car, R.; Lee, C.; Vanderbilt, D. *Phys. Rev. B* **1993**, *47* (16), 10142.

(66) Saukkoriipi, J. J.; Laasonen, K. *J. Phys. Chem. A* **2008**, *112* (43), 10873.

(67) Zaworotko, M. J.; Rogers, R. D.; Atwood, J. L. *Organometallics* **1982**, *1*, 1179.

(68) Shreve, A. P.; Mulhaupt, R.; Fultz, W.; Calabrese, J.; Robbins, W.; Ittel, S. D. *Organometallics* **1988**, *7*, 409.

(69) Fratiello, A.; Lee, R. E.; Nishida, V. M.; Schuster, R. E. *J. Chem. Phys.* **1968**, *48* (8), 3705.

(70) Swaddle, T. W.; Rosenqvist, J.; Yu, P.; Bylaska, E.; Phillips, B. L.; Casey, W. H. *Science* **2005**, *308*, 1450.

(71) Casey, W. H. *Chem. Rev.* **2006**, *106* (1), 1.

(72) Johansson, G.; Lundgren, G.; Sillen, L. G.; Söderquist, R. *Acta Chem. Scand.* **1960**, *14*, 769.

(73) Johansson, G. *Acta Chem. Scand.* **1960**, *14* (3), 771.

(74) Keggin, J. F. *Proc. R. Soc., Ser. A* **1934**, *144*, 75.

(75) Lin, I.-C.; Seitsonen, A. P.; Coutinho-Neto, M. D.; Tavernelli, I.; Rothlisberger, U. *J. Phys. Chem. B* **2009**, *113* (4), 1127.

(76) Harlan, C. J.; Mason, M. R.; Barron, A. R. *Organometallics* **1994**, *13* (8), 2957.

(77) Soper, A. K.; Bruni, F.; Ricci, M. A. *J. Chem. Phys.* **1997**, *106* (1), 247.

(78) Soper, A. K. *J. Phys.: Condens. Matter* **2007**, *19* (33), 335206.

(79) Kuo, I.-F. W.; Mundy, C. J.; McGrath, M. J.; Siepmann, J. I.; VandeVondele, J.; Sprik, M.; Hutter, J.; Chen, B.; Klein, M. L.; Mohamed, F.; Krack, M.; Parrinello, M. *J. Phys. Chem. B* **2004**, *108* (34), 12990.

(80) VandeVondele, J.; Mohamed, F.; Krack, M.; Hutter, J.; Sprik, M.; Parrinello, M. *J. Chem. Phys.* **2005**, *122*, 014515-1.

(81) Lee, H.-S.; Tuckerman, M. E. *J. Chem. Phys.* **2007**, *126*, 164501.

(82) Tuckerman, M.; Laasonen, K.; Sprik, M.; Parrinello, M. *J. Chem. Phys.* **1995**, *103* (1), 150.

(83) Tuckerman, M. E.; Ungar, P. J.; von Rosenvinge, T.; Klein, M. L. *J. Phys. Chem.* **1996**, *100* (31), 12878.

CT900670A

JCTC Journal of Chemical Theory and Computation

# Potential of Mean Force Calculations: A Multiple-Walker Adaptive Biasing Force Approach

K. Minoukadeh,[*,†,‡] C. Chipot,[§,||] and T. Lelièvre[†,‡]

*CERMICS, École des Ponts ParisTech, 6−8 avenue Blaise-Pascal, 77455 Champs-sur-Marne, Marne-la-Vallée cedex 2, France, MICMAC Project-Team, INRIA Rocquencourt, 78153 Le Chesnay, France, Équipe de dynamique des assemblages membranaires, UMR 7565, Nancy Université, BP 239, 54506 Vandœuvre-lès-nancy Cedex, France, and Theoretical and Computational Biophysics Group, Beckman Institute, University of Illinois at Urbana−Champaign, Urbana, Illinois 61801*

Received October 5, 2009

**Abstract:** The adaptive biasing force (ABF) scheme is a powerful molecular-dynamics based method for overcoming barriers of the free-energy landscape. Integration of the mean force measured along a chosen reaction coordinate (RC) yields the so-called potential of mean force (PMF). The RC is a coarse-grained description of the transition mechanism. The mean force is estimated by accruing and averaging the instantaneous force exerted on the system. The PMF is then used to bias the standard dynamics of the system in order to improve sampling in the RC. We show that faster exploration of the reaction pathway can be achieved by running multiple walkers in parallel and exchanging information at fixed intervals in the course of the simulation. Numerical experiments performed on the prototypical deca-alanine peptide demonstrate that the convergence properties of the free-energy calculation are globally improved through a more efficient exploration of compact configurations reflected in parallel valleys of the free-energy landscape. Diffusion along the RC is further enhanced by a selection mechanism, whereby far-reaching walkers are cloned, replacing less effective ones.

## 1. Introduction

Central to the understanding of most processes of either physical, chemical, or biological interest, the determination of the underlying free-energy change occupies a prominent position in the arena of numerical simulations. Over the past decades, a variety of methods have been devised to compute free-energy differences efficiently (see, for example, refs 1 and 2). Roughly speaking, these methods can be classified into two main categories: (i) the free energy is computed directly, or (ii) its first derivative is determined and subsequently integrated. Perturbation techniques,[3] probability density function-based methods such as histogram methods,[4−6] nonequilibrium computations,[7] and adaptive biasing potential methods,[8,9] for example, fall into the first category. Thermodynamic integration[10] and adaptive biasing force methods,[11,12] which are the core of the present work, belong to the second category. Adaptive methods are designed to compute free-energy profiles and favor transitions between metastable states by using a current estimate of the free energy as a biasing potential.

In this contribution, we are interested in a particular class of adaptive methods, referred to as adaptive biasing force methods.[11−13] Specifically, we endeavor to investigate a novel implementation of this class of methods, using a number of walkers simulated in parallel, in the spirit of the ideas put forth by Lelièvre et al.[14] The advantage of the present, novel implementation is 3-fold. First, the parallelization is straightforward, and its theoretical parallel efficiency is very good since the only shared information is the biasing force, or the marginal law, namely, low-

[*] To whom correspondence should be addressed: E-mail: kimiya.minoukadeh@cermics.enpc.fr.

[†] École des Ponts ParisTech.

[‡] INRIA Rocquencourt.

[§] Nancy Université.

[||] University of Illinois at Urbana−Champaign.

10.1021/ct900524t
Published on Web 03/04/2010

dimensional functions. In turn, this yields efficient, scalable algorithms to compute free-energy differences, well adapted to the massively parallel architecture of high-performance computers. Second, we will show that the implementation relying upon many walkers is particularly interesting when the reaction coordinate does not describe well all the metastabilities of the system, which, quite unfortunately, constitutes a generic situation for the vast majority of nontrivial molecular systems. This is typically the case, for instance, of the so-called bichannel scenario—namely, the free-energy landscape features two parallel valleys, which are orthogonal to the isocontours of the reaction coordinate—or, more generally, when several transition mechanisms are associated with a single reaction coordinate, which is, therefore, not sufficient to parametrize fully the transformation. The underlying idea is that, when many walkers are involved, they can visit more efficiently in parallel all the valleys in the direction of the reaction coordinate. A mathematical proof is currently underway to show that, in the limiting case of a very large number of walkers, and with suitable assumptions, the rate of convergence of the ABF method is in fact not limited by free-energy barriers orthogonal to the RC direction. Third, as will be detailed below, this new implementation allows selection mechanisms to be introduced, consisting of duplicating effective walkers, while deleting poor ones, according to a fitness function that ought to be chosen. An example of such a fitness function, which favors rapid exploration of the reaction coordinate, will be provided hereafter.

As a proof of concept, the present approach was probed on a realistic test case, using a high-level Tcl implementation of the algorithm in the scalable molecular dynamics program NAMD.[15−17] The efficiency of the overall procedure is, however, expected to be enhanced by embedding and optimizing the algorithm at a deeper level of the molecular-dynamics platform.

In the following section, the mathematical framework of the method is introduced and the adaptive biasing force method reviewed. Next, the discretization and implementation details are presented. The present contribution closes with a discussion of the numerical results obtained for the reversible folding of the paradigmatic deca-alanine peptide.

## 2. General Setting

In the canonical ensemble, a system of dimension $d$ is equipped with the Boltzmann−Gibbs probability measure, i.e., the canonical measure:

$$\mu(dq) = \phi(q)dq = Z^{-1}\exp(-\beta V(q))\,\mathrm{d}q \qquad (1)$$

where $\phi$ is the density of the measure, $\beta = 1/(k_{\mathrm{B}}T)$ is proportional to the inverse temperature, $q \in \mathbb{R}^d$ is the system configuration, $V : \mathbb{R}^d \to \mathbb{R}$ is the potential energy function, and $Z = \int_{\mathbb{R}^d} \exp(-\beta V(q))\,\mathrm{d}q$ is the normalization constant or the so-called partition function. To sample this measure, one can use the overdamped Langevin dynamics:[18]

$$dX_t = -\nabla V(X_t)dt + \sqrt{2\beta^{-1}}\,\mathrm{d}W_t \qquad (2)$$

where $(X_t)_{t\geq 0}$ is the system trajectory and $W_t$ is an $\mathbb{R}^d$-valued standard Brownian motion (or Wiener process). Under suitable regularity assumptions on the potential, the dynamics (eq 2) is ergodic and admits the canonical measure as its unique invariant measure. It must be emphasized that, for the sake of simplicity, the method is described in the framework of the overdamped dynamics. The method can, nevertheless, be generalized to the Langevin dynamics as is done in the numerical simulations at the end of the paper.

The canonical measure (eq 1) gives us microscopic information about the system, the probability that it is to be found at any particular point $q$ in configuration space. A practitioner, however, is generally interested in some coarse-grained collective variable $\xi(q)$, where $\xi$ is typically a smooth mapping from $\mathbb{R}^d$ to $\mathbb{R}$. In what follows, $\xi$ will be referred to as the reaction coordinate (RC). The RC typically represents an end-to-end distance of a protein chain, a structural angle in a protein, or a measure of the evolution of a chemical system. If $X$ is a random variable with probability $\mu$, then $\xi(X)$ is a random variable with a law whose density $\phi^{\xi}$ is defined by

$$\phi^{\xi}(z) = \int_{\mathbb{R}^d} \phi(q)\,\delta(\xi(q) - z)\,\mathrm{d}q \qquad (3)$$

and the distribution $\phi^{\xi}(z)dz$ is called the marginal distribution of $\mu$ in $\xi$. The free energy, or so-called potential of mean force (PMF), $A$, is related to this marginal density in the following way

$$A(z) = -\beta^{-1}\ln \phi^{\xi}(z) \qquad (4)$$

Sampling the canonical measure using the standard overdamped Langevin dynamics (eq 2) can in fact be inefficient in practice. The convergence to equilibrium can be very slow due to metastable states where the dynamics remains trapped for long periods of time. To explore the whole configuration space, one often needs to overcome very large energy barriers. From Arrhenius's law, it follows that the typical time needed to overcome these barriers scales exponentially with the barrier height. Regular molecular-dynamics methods are, therefore, typically not used to calculate statistical averages for systems prone to metastabilities.

Several methods have been proposed to ameliorate sampling methods in these situations, such as the *blue moon* method[19] or importance sampling methods such as *umbrella sampling*.[20] More recently, adaptive importance sampling methods have been developed such as the *Wang−Landau* method[8] and the *adaptive biasing force* (ABF) method.[11] The latter and its variations will be the focus of this contribution.

Before detailing the ABF method, the reader is reminded that the main quantity of interest in the study of chemical reactions is a free-energy difference and not an absolute free energy. Free energies are, therefore, computed only up to an additive constant. The free-energy difference between two coarse-grained states, labeled by the RC values $z_a$ and $z_b$, can be written as

$$\Delta A = A(z_b) - A(z_a) = \int_{z_a}^{z_b} A'(z)\,\mathrm{d}z \qquad (5)$$

where ′ is the derivative with respect to the collective variable value $z$ and $A'$ is called the mean force. The integrand can be shown to be the Boltzmann average of a real-valued function $F^V$, conditioned to being at a fixed point $z$ in the reaction coordinate space

$$\begin{aligned} A'(z) &= \frac{\int_{\mathbb{R}^d} F^V(q)\exp(-\beta V(q))\,\delta(\xi(q)-z)\,\mathrm{d}q}{\int_{\mathbb{R}^d}\exp(-\beta V(q))\,\delta(\xi(q)-z)\,\mathrm{d}q} \\ &= \langle F^V(q)|\xi(q)=z\rangle_\mu \end{aligned} \qquad (6)$$

where

$$F^V = \frac{\nabla V\cdot\nabla\xi}{|\nabla\xi|^2} - \beta^{-1}\nabla\cdot\left(\frac{\nabla\xi}{|\nabla\xi|^2}\right) \qquad (7)$$

and $\langle\cdot\rangle_\mu$ represents the canonical average—i.e., the average with respect to the measure $\mu$. Note that $F^V$ is the negative projection of the force onto the RC plus some correction term. For the derivation of eq 7, the reader is referred to refs 21, 22, and 23. The aim of the ABF method, which will be detailed hereafter, is to compute $A'$ as efficiently as possible.

## 3. Adaptive Biasing Force Methods

In this section, we will present the framework behind ABF methods for free energy computations.

**3.1. Framework.** The basic idea of ABF is to use the mean force estimate to bias the dynamics and help the system overcome free-energy barriers. An estimate of $A'(z)$ is obtained as the statistical average of the force field $F^V$ at specified points $z$ along the RC by accruing instantaneous forces $F^V(X_t)$ for a single system trajectory $X_t$ when $\xi(X_t) = z$. In the long-time limit, one obtains a good approximation for $A'$, and $\Delta A$ can be computed by numerical integration. The resulting biased dynamics is

$$\begin{cases} dX_t = -\nabla(V - A_t\circ\xi)(X_t)\,\mathrm{d}t + \sqrt{2\beta^{-1}}\,\mathrm{d}W_t \\ A_t'(z) = \langle F^V(X_t)|\xi(X_t)=z\rangle \end{cases} \qquad (8)$$

where $A_t\circ\xi$ denotes the composition of $A_t$ with $\xi$, so that $A_t\circ\xi(x) = A_t(\xi(x))$, and $A_t'$ is the estimated mean force. [Note that the gradient term in the biased dynamics can be rewritten as $-\nabla V + (A_t'\circ\xi)\nabla\xi$; thus, only estimated mean force information is needed and not the estimated free energy.] The estimated mean force is thus defined as a conditional average of $F^V(X_t)$ at a fixed value of $\xi(X_t)$. In practice, it can be approximated as an average over many walkers or as an *on-the-fly* average over the trajectory $X_t$ (see the next section for more details). The above can be viewed as an overdamped Langevin dynamics, with the potential $V$ replaced by the time varying potential $\mathcal{V}_t = V - A_t\circ\xi$. In the following, $\psi_t$ will denote the density of the law of $X_t$. The consistency of the method may be justified by noticing that, if a stationary state $A_\infty'$ for $A_t'$ is obtained, then $\psi_\infty$ is proportional to $\exp(-\beta\mathcal{V}_\infty)$, and thus $A_\infty' = A'$ since

$$\begin{aligned} A_\infty'(z) &= \langle F^V(q)|\xi(q)=z\rangle_{\exp(-\beta\mathcal{V}_\infty)} \\ &= \langle F^V(q)|\xi(q)=z\rangle_\mu = A'(z) \end{aligned}$$

As a result, $A_t$ converges to $A$ up to an additive constant, and the equilibrium marginal density in $\xi$ is constant, since

$$\langle\delta(\xi(q)-z)\rangle_{\exp(-\beta\mathcal{V}_\infty)} = \frac{\langle\delta(\xi(q)-z)\rangle_\mu}{\exp(-\beta A(z))}$$

is constant by the definitions of $\mu$ and $A$ in eqs 1 and 4, respectively. Precise convergence results can be found in ref 24. The aim of ABF is, therefore, to estimate the biasing force as efficiently as possible in order to bias the dynamics by reducing and eventually eliminating any force along $\xi$. It serves as an adaptive importance sampling method, driving the system out of its metastable states, using on-the-fly estimates of the mean force.

**3.2. Calculating the Bias.** Different approaches have been proposed in recent literature[11,24,25] for calculating the biasing force. There are two principal methods for computing $A_t'$, which will serve as a basis of comparison in the present contribution.

*Original ABF.* The idea of the standard ABF method,[11] which involves one single walker, is to calculate averages using the whole trajectory of the system. The mean force is calculated by taking a trajectorial average of instantaneous forces at fixed $z$ using one long system trajectory (see ref 1 for further details)

$$\langle F^V(X_t)|\xi(X_t)=z\rangle \simeq \frac{\int_0^t F^V(X_s)\,\delta(\xi(X_s)-z)\,\mathrm{d}s}{\int_0^t \delta(\xi(X_s)-z)\,\mathrm{d}s} \qquad (9)$$

The mean force estimate is only computed once a trajectory reaches the value $z$ in the RC; therefore, the denominator in the above equation is always nonzero for the RC values needed.

*Multiple Walker (MW-)ABF.* In a recent paper,[25] a new formulation of the ABF method has been proposed, consisting of running $R > 1$ trajectories of the ABF dynamics in parallel. The $R$ walkers of the system follow a similar dynamics driven by independent Brownian motions. These multiple walkers then exchange information at fixed time intervals. The immediate gain of this new formulation is that one can take advantage of parallel computing to speed up convergence of the ABF method. Furthermore, with the use of a small number of walkers, we are able to overcome issues related to poorly chosen or oversimplified reaction coordinates, where other important slow degrees of freedom are overlooked. In such cases, metastabilities can be found at fixed $\xi$, as illustrated in Figure 1. With multiple walkers, it is likely that each walker will explore a different path or valley along $\xi$, whereas single-walker simulations could potentially take exponentially long times to fully explore the low energy states. This will be studied numerically in the final sections of the paper.

In the following, $(X_t^i)_{0\le i\le R-1}$ is the set of trajectories for the $R$ walkers. Each trajectory $X_t^i$ follows the dynamics (eq 8) with the Brownian motion $W_t$ replaced with $W_t^i$. The biasing force is then estimated using an average over all the trajectories and over all walkers:

$$\langle F^V(X_t)|\xi(X_t) = z\rangle \simeq \frac{\sum_{i=0}^{R-1} \int_0^t F^V(X_s^i)\, \delta(\xi(X_s^i) - z)\, \mathrm{d}s}{\sum_{i=0}^{R-1} \int_0^t \delta(\xi(X_s^i) - z)\, \mathrm{d}s} \tag{10}$$

Implementation details for the approaches discussed above will be discussed in the next section.

**3.3. Enhancing Sampling through Selection.** In addition to the exchange of information between walkers to compute the estimated mean force $A_t'$, the MW-ABF method allows for resampling of the walkers according to their "importance". The success of the ABF method is strongly determined by the marginal distribution of walkers in the RC, given that the RC has been well chosen. One would, therefore, want to encourage walkers that are exploring undersampled regions of the RC and penalize those in oversampled regions. A selection mechanism[25-27] may be used to achieve this objective. It is implemented by a system of interacting walkers, where the walkers are cloned or killed at a rate defined by $S(t, z)$ over the values taken by the RC. The function $S(t, z)$ can be chosen as

$$S(t, z) = c\frac{\partial_{zz}\psi_t^\xi(z)}{\psi_t^\xi(z)} \tag{11}$$

where $c$ is a positive constant and $\psi_t^\xi$, defined by

$$\psi_t^\xi(z) = \int_{\mathbb{R}^d} \psi_t(q)\, \delta(\xi(q) - z)\, \mathrm{d}q$$

represents the marginal distribution of walkers in the RC at time $t$. Here, $\psi_t(q)$ denotes the distribution of $X_t$. With this choice of the function $S$, it can be shown that the marginal density $\psi^\xi$ satisfies the partial differential equation (in fact for a slightly modified version of the original adaptive dynamics (eq 8), see ref 25):

$$\partial_t\psi_t^\xi = (\beta^{-1} + c)\partial_{zz}\psi_t^\xi \tag{12}$$

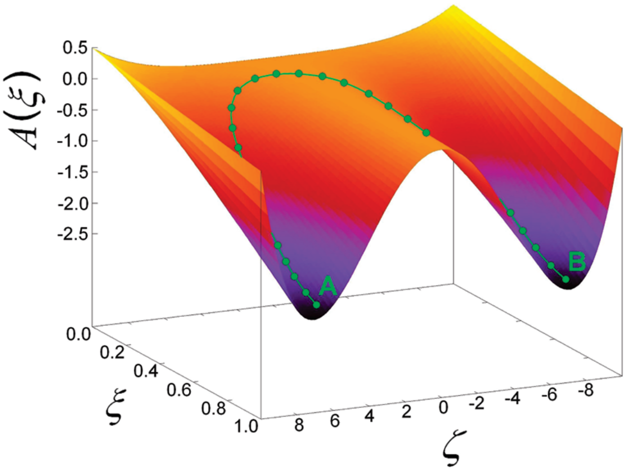


***Figure 1.*** Example of a 2-dimensional free-energy surface exhibiting metastabilities at fixed $\xi$. The variable $\zeta$ represents another slow degree of freedom of the system, orthogonal to the RC. The standard ABF method relies on efficient sampling at fixed points in the RC, which is made difficult by the presence of such large energy barriers in the orthogonal directions. Using multiple walkers helps to overcome this issue as each one is very likely to explore a different pathway.

The selection process thus accelerates the diffusion of the marginal distribution in the RC. The reason for this choice of $S$ can also be understood when written in a finite difference form

$$S(t, z) \simeq \frac{3c}{\Delta z^2\psi_t^\xi(z)}\left[\frac{\psi_t^\xi(z - \Delta z) + \psi_t^\xi(z) + \psi_t^\xi(z + \Delta z)}{3} - \psi_t^\xi(z)\right] \tag{13}$$

where $\Delta z$ is some small displacement in the RC. The quantity $S(t, z)$ at a given value $z$ of the RC is, therefore, positive if the marginal density at this point is small compared to its local average, and negative otherwise. To implement the selection process, one can either continuously update birth and death times, initially drawn from an exponential distribution, as in ref 25, or resample the walkers according to their weights at fixed resampling times. The latter will be used for the simulations reported herein. At a resampling time $t$, each walker trajectory $X_t^i$ is given a weight:[28]

$$w_t^i = K_t^{-1} \exp[\int_0^t S(s, \xi\{X_s^i\})\, \mathrm{d}s] \tag{14}$$

where $K_t = \Sigma_{i=0}^{R-1} \exp[\int_0^t S(s, \xi\{X_s^i\})\, \mathrm{d}s]$ is the normalization constant. Replicas are initially assigned a uniform weight, $w_0^i = 1/R$, which evolves in time. From eqs 13 and 14, it is now clear that a walker $i$ that is often found in undersampled regions—in which case $S(t, \xi(X_t^i))$ is often positive—is given a stronger weight than walkers in oversampled regions—where $S(t, \xi(X_t^i))$ is often negative. The $i$th walker is then replicated on average $Rw_t^i$ times. This procedure thus accelerates the convergence to a uniform distribution of the walkers in the RC, in accordance with eq 12.

Let us now give some details about the resampling procedure. To calculate the number of times a walker is to be copied, a systematic resampling method[29-31] is used, described briefly by the following algorithm. At a resampling time $t$:

```
Set $u \sim U(0, 1)$, $\bar{w}_0 = w_t^0$, $N_0 = \lfloor R \times \bar{w}_0 + u \rfloor$,
for $i = 1, ..., R - 1$
    $\bar{w}_i = \bar{w}_{i-1} + w_t^i$,
    $N_i = \lfloor R \times \bar{w}_i + u \rfloor - \lfloor R \times \bar{w}_{i-1} + u \rfloor$
end
```

where $U(0, 1)$ denotes a uniform distribution between 0 and 1, $w_t^i$ is the normalized weight assigned to walker $i$ as defined in eq 14, $\bar{w}_i$ is the cumulative sum of the weights, $\lfloor \cdot \rfloor$ is the integer part, and $N_i$ is the number of copies of walker $i$ to be generated. It is important to note that this algorithm guarantees that $\Sigma_{i=0}^{R-1} N_i = R$. After every resampling stage, the weights of all walkers are reset uniformly to the value $1/R$. The choice of the constant $c$ in eq 11 is of paramount importance in the performance of the selection mechanism. This parameter should

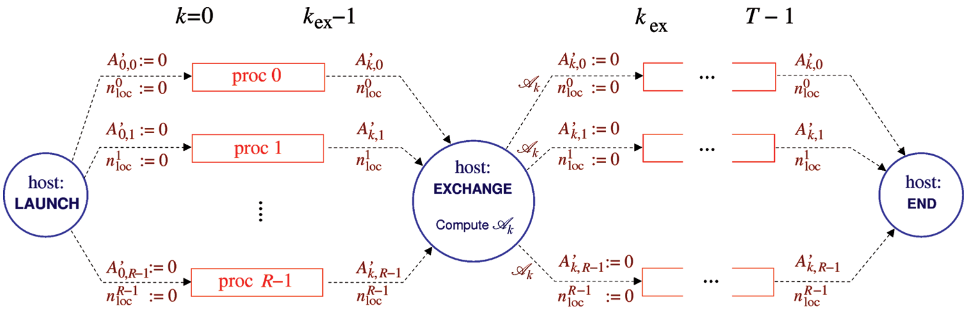


***Figure 2.*** Schematic diagram of MW-ABF. The main script is executed on a host machine, which acts as the Tcl server. This machine launches the $R$ walkers onto different processors via socket connections and, after every $k_{ex}$ time steps, carries out exchange of information. This consists of reading in local variables from each processor; computing the total biasing force $A_k'$ by means of eqs 21 and 22, sharing $\mathscr{A}_k'$ with all processors, and setting local variables to zero. This is carried out $T/k_{ex}$ times until the program terminates.

be sufficiently large to accelerate the exploration along the RC, but not too large in case one walker is selected during the resampling stage (due to degeneracy of weights), which implies a very large variance of the estimator. This will be discussed further at the end of the next section.

## 4. Implementation Details

In this section, the implementation details of the adaptive biasing force methods are provided. The simulations reported in the present contribution have been carried out using the scalable molecular-dynamics code NAMD, but the algorithmic detail is by no means specific to this software package. The ABF methods have been implemented as Tcl scripts, for which the single-walker ABF method is already available. How the method is discretized will be spelled out hereafter, and the detail of the single-walker ABF method will be outlined before proceeding with the implementation of the MW-ABF method and selection.

We consider a reaction coordinate $\xi$ taking values in the interval $[z_0, z_N]$, which is divided into $N$ bins of size $\Delta z = (z_N - z_0)/N$. We denote by $\tilde{\xi}:\mathbb{R}^d \to \{0, ..., N-1\}$ a mapping from a configuration onto its associated bin in the RC

$$\tilde{\xi}(\cdot) = \left\lfloor \frac{\xi(\cdot) - z_0}{\Delta z} \right\rfloor$$

where $\lfloor \cdot \rfloor$ again denotes the integer part. In the following, functions and trajectories will be indexed by the number of time steps $k$, so that $A_k'$ will be the mean force approximation and $X_k$ will be the configuration of the system at time $k\Delta t$, for a time step $\Delta t$. Furthermore, with a slight abuse of notation, $z$ will now denote the bin in the reaction coordinate, $z = \tilde{\xi}(X_t)$.

**Original ABF Method.** The reader is reminded that, in the standard ABF method, the biasing force is calculated for each bin using a trajectorial average, as in eq 9. The biasing force is in practice updated to include the current force observation. For $z \in \{0, ..., N-1\}$

$$A_k'(z) = \frac{n_{\text{tot}}(k-1, z)}{n_{\text{tot}}(k, z)} A_{k-1}'(z) + \frac{\mathbf{1}_{\tilde{\xi}(X_{k-1})=z}}{n_{\text{tot}}(k, z)} F^V(X_{k-1}) \tag{15}$$

where $\mathbf{1}_{\tilde{\xi}(X_k)=z}$ denotes the indicator function—taking value 1 if $\tilde{\xi}(X_k) = z$ and 0 otherwise—and

$$n_{\text{tot}}(k, z) = \sum_{s=0}^{k-1} \mathbf{1}_{\tilde{\xi}(X_s)=z} \tag{16}$$

is the total number of times the system trajectory has visited bin $z$. To justify eq 15, the expression in eq 9 is recast in its discretized form

$$A_k'(z) = \frac{\sum_{s=0}^{k-1} F^V(X_s)\mathbf{1}_{\tilde{\xi}(X_s)=z}}{\sum_{s=0}^{k-1} \mathbf{1}_{\tilde{\xi}(X_s)=z}} \tag{17}$$

Developing further, one subsequently obtains

$$\begin{aligned} A_k'(z) &= \frac{\sum_{s=0}^{k-2} F^V(X_s)\mathbf{1}_{\tilde{\xi}(X_s)=z} + F^V(X_{k-1})\mathbf{1}_{\tilde{\xi}(X_{k-1})=z}}{n_{\text{tot}}(k, z)} \\ &= \frac{n_{\text{tot}}(k-1, z)A_{k-1}'(Z) + F^V(X_{k-1})\mathbf{1}_{\tilde{\xi}(X_{k-1})=z}}{n_{\text{tot}}(k, z)} \end{aligned}$$

where the last line follows from eq 17 at time $k-1$.

**MW-ABF.** The basis for the multiple-walker implementation of ABF in NAMD can be found in a set of Tcl scripts written for parallel-tempering, replica-exchange simulations.[15,17] The scripts use Tcl server and socket connections to launch NAMD processes for each individual walker. Each walker is handled by a different computing unit. The NAMD processes run for a fixed number of time steps, then wait in order for the Tcl server to exchange information between walkers. Figure 2 is a synoptic diagram of the MW-ABF method. It is not necessary (and not desirable from a computational point of view) to exchange information at every time step. Exchange of information between walkers only occurs at every $k_{ex}$ time steps (Figure 3). We therefore proceed as follows: the mean force approximation, denoted by $A_{k,i}'(\bullet)$, is evaluated locally on the computing unit, where the indices $k$ and $i$ represent respectively the number of time steps since the beginning of the simulation and the

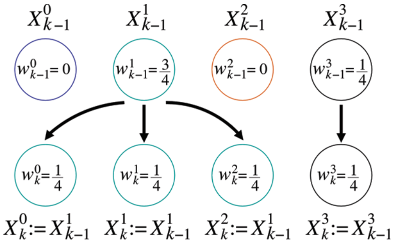


***Figure 3.*** Selection mechanism for $R = 4$ walkers. If walker $i$ has weight $w^i_k$ at the time of selection, on average, $Rw^i_k$ copies are made of this walker at the next step. In practice, this means that $Rw^i_k$ walkers will be launched using the configuration and velocity files of walker $i$. Note that, in the above, $k = nk_{ex}$, the time at which selection is carried out.

computing unit running walker $i$. This quantity therefore depends solely on the trajectory of the walker of interest. Between exchange times, $k \in [nk_{ex}, (n + 1)k_{ex}]$, the mean force estimation evolves according to the update formula

$$A'_{k,i}(z) = \frac{n^i_{loc}(k-1,z)}{n^i_{loc}(k,z)} A'_{k-1,i}(z) + \frac{\mathbf{1}_{\tilde{\xi}(X^i_{k-1})=z}}{n^i_{loc}(k,z)} F^V(X^i_{k-1}) \tag{18}$$

Note that this is the same as eq 15, where $A'_k$ is replaced by $A'_{k,i}$, $X_k$ is replaced by $X^i_k$, and $n_{tot}(k, z)$ is replaced by $n^i_{loc}(k, z)$, the number of times walker $i$ has entered bin $z$ since the last exchange, defined in eq 19 below.

At every exchange time, the information gathered by each walker is collected and local variables on each processor are updated. This is formalized with the help of some further notation. We denote by

$$k_{last} = \lfloor k/k_{ex} \rfloor k_{ex}$$

the time of the last exchange. Next,

$$n^i_{loc}(k,z) = \sum_{s=k_{last}}^{k-1} \mathbf{1}_{\tilde{\xi}(X^i_s)=z} \tag{19}$$

denotes the number of times walker $i$ has entered bin $z$ since the last exchange, and

$$n^i_{tot}(k,z) = \sum_{s=0}^{k-1} \mathbf{1}_{\tilde{\xi}(X^i_s)=z} \tag{20}$$

is the total number of times walker $i$ has entered bin $z$ since the beginning of the simulation. Finally,

$$N_{loc}(k,z) = \sum_{i=0}^{R-1} n^i_{loc}(k,z) \text{ and } N_{tot}(k,z) = \sum_{i=0}^{R-1} n^i_{tot}(k,z)$$

denote respectively the total number of visits to bin $z$ since the last exchange and the beginning of the simulation, over all the walkers.

At every exchange time, a local average $A'_{loc}$ is calculated of the mean force estimated from the run of each individual walker:

$$A'_{loc}(k,z) = \frac{1}{N_{loc}(k,z)} \sum_{i=0}^{R-1} n^i_{loc}(k,z)\, A'_{k,i}(z) \tag{21}$$

The total biasing force, $\mathcal{A}'_k(z)$, to be shared between the walkers, is then also updated to include this new information

$$\mathcal{A}'_k(z) = \left[1 - \frac{N_{loc}(k,z)}{N_{tot}(k,z)}\right] \mathcal{A}'_{k-1}(z) + \frac{N_{loc}(k,z)}{N_{tot}(k,z)} A'_{loc}(k,z) \tag{22}$$

The latter quantity, $\mathcal{A}'_k$, is then communicated to each one of the walkers, and the variables $n^i_{loc}$ and $A'_{k,i}$ are reset to zero.

The total biasing force in eq 22 is utilized by each walker in the steps following the exchange; however, new local information is also incorporated in order to speed up the diffusion in $\xi$. The biasing force applied to walker $i$ in the simulations, therefore, writes

$$F^i_{bias,k}(z) = \left[1 - \frac{n^i_{loc}(k,z)}{N_{tot}(k,z)}\right] \mathcal{A}'_{k_{last}}(z) + \frac{n^i_{loc}(k,z)}{N_{tot}(k,z)} A'_{k,i}(z)$$

where $\mathcal{A}'_{k_{last}}$ is again the total mean force calculated at the preceding exchange interval, and $A'_{k,i}$ is the local mean force information, as defined in eq 18. [In practice, this force is actually only fully applied to the dynamics after a certain number of visits have been made to the bin, that is to say, after $N_{tot}(k, z) > N_{min}$, where in our simulations $N_{min} = 500$. If $N_{tot}(k, z) < N_{min}/2$, then no biasing force is added. Beyond that, the force is slowly introduced using a ramp function with scaling factor $\min(2N_{tot}/N_{min} - 1, 1)$.]

**Selection.** Resampling may be carried out at most every $k_{ex}$ time steps, when the walkers exchange information. Selection is a technically costly process as NAMD must be exited and reloaded with new configuration and velocity files. For this reason, it is even advisible for it to be carried out less frequently. The computational complexity of the process is $O(R)$ using a systematic resampling method (see the previous section for the algorithm). For purposes of illustration, the resampling will be carried out as often as the interprocessor communication, namely, every $k_{ex}$ time steps. The purpose of resampling is to improve the exploration in the RC. The weights of the walkers are adjusted according to the utility function $S(k, z)$, depending in practice upon the total distribution of the walkers:

$$S(k,z) = c\,\frac{N_{tot}(k_{last}, z+1) - 2N_{tot}(k_{last}, z) + N_{tot}(k_{last}, z-1)}{N_{tot}(k_{last}, z)} \tag{23}$$

The integral in eq 14 is calculated by summing the terms $S(k, \tilde{\xi}(X^i_k))$ over $k$ for each walker $i$ during each individual run. At the selection stage, when $k = k_{ex}$, the weights of the walkers are computed:

$$w^i_k = K_k^{-1} \exp[\sum_{s=k_{last}}^{k-1} S(s, \tilde{\xi}\{X^i_s\})]$$

where $K_k$ is again the normalization constant. The walkers are then selected according to these weights using a systematic resampling method, as described above. In practice, to generate $N_i$ copies of walker $i$, the configuration and velocity files for walker $i$ are passed to NAMD as the startup files for $N_i$ walkers. Finally, after resampling, $S$ is set to zero, so that all walker trajectories have equal weight.

For resampling to be effective, there are two main issues that need to be addressed. First, the constant $c$ has to be chosen carefully: it must be large enough for the selection mechanism to be beneficial and small enough to avoid degeneration of weights, where all walkers are given zero weight except for one. Another issue to be addressed is when to stop resampling. Due to the technical costs of the selection mechanism, it is advised to impose a stopping criterion, so that selection is not applied throughout the whole simulation. The stopping criterion could depend on the sampling of the RC, or on the distribution of the weights. For the simulations herein, the latter criterion is used, for once the walkers begin to be equally weighted, the selection has no effect and ought to be stopped. In order to measure the distribution of the weights, we consider the relative entropy of the weights compared to a uniform distribution, defined by

$$E_w(t) = \sum_{i=0}^{R-1} w_i \log(Rw_i) \qquad (24)$$

This can be understood as the difference between the entropy of weights and the entropy for the uniform weight distribution: $\sum_{i=0}^{R-1} w_i \log(w_i) - \log(1/R)$. This quantity is bounded above by $\log(R)$, in case of degeneracy, and is bounded below by 0, in case of uniform distribution of weights. A good stopping criterion for the selection algorithm is to end the process when the relative entropy is sufficiently small. In our simulations, the selection is stopped once

$$E_w(t) < \varepsilon \log(R) \qquad (25)$$

where $0 < \varepsilon < 1$ is set closer to 0 for a stringent stopping criterion or closer to 1 for a weaker threshold.

## 5. Numerical Results

In this section, we present comparisons of the single-walker and multiple-walker ABF methods on the deca-alanine peptide in the gas phase, for which comprehensive studies have already been carried out.[12,32,33] All the simulations reported herein were performed with the molecular-dynamics code NAMD,[15−17] using the CHARMM27[34] force field. The 10-residue peptide chain has a total of 104 atoms and the RC has been chosen as the distance separating the center of mass of its first and last C−H pairs. To sample the full range of conformations from the α-helical conformation to the ensemble of extended structures, the range of values accessible to $\xi$ varies from 12 to 32 Å. Additional tests are also carried out to study more compact conformations, where $\xi$ varies between 4 and 16 Å. The system is kept within the assigned ranges by enforcing reflective boundary conditions.

The average forces were accumulated in bins of size $\Delta z = 0.1$ Å. The equations of motion were integrated employing Langevin dynamics with a time step $\Delta t = 0.5$ fs.

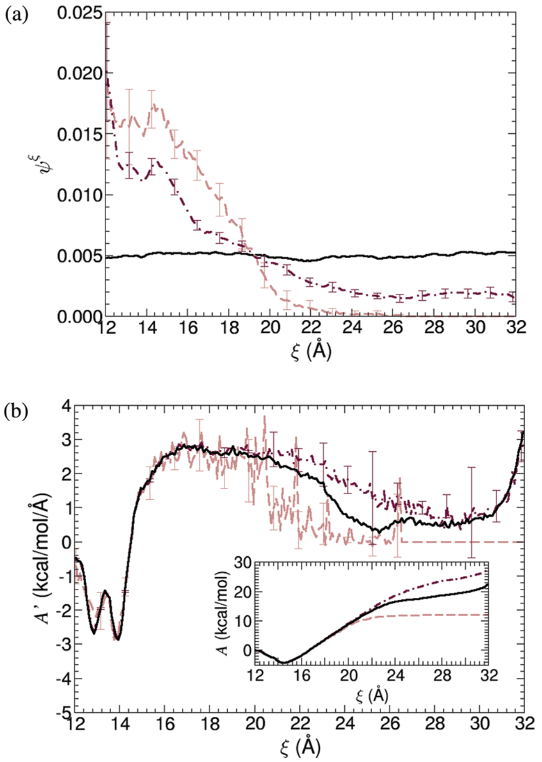


***Figure 4.*** Results for $\xi$ ranging from 12 to 32 Å (after 0.25 ns). The curves are averages of 20 independent single-walker (dashed lines) and 16-walker (dashed-dotted lines) simulations with error bars representing the 95% confidence intervals. Solid lines represent reference profiles, obtained from a single-walker run of 200 ns. (a) Density of marginal distribution in the RC. The multiple-walker simulation has explored the whole $\xi$-space whereas the single-walker simulations very rarely stretch beyond 22 Å. (b) Mean force and free-energy profiles (inset). For the multiple-walker simulations, we see the mean force profiles already nearly converged, whereas little information is gathered beyond 22 Å for the standard ABF simulation.

Electrostatic and van der Waals interactions were truncated smoothly beyond 11 Å.

We will first present the results for the conventional range of 12 to 32 Å, which spans conformations comprised between the α-helix to more extended structures. Next, results for more compact conformations—with $\xi$ ranging from 4 to 16 Å—are presented, where stark differences can be observed between the single- and multiple-walker ABF simulations. Finally, we will study the impact of selection on walkers.

**5.1. Reaction Coordinate Range: 12−32 Å.** Starting from the α-helical conformation, $R$ walkers of the system are launched with ABF dynamics, communicating at every $k_{ex} = 50\,000$ time steps (25 ps). Reference curves are obtained from a 200-ns simulation using the original ABF algorithm, featuring a single walker.

Figure 4 compares the sampling distribution, mean force, and free-energy profiles for single-walker and 16-walker simulations after 0.25 ns.

It may be observed from Figure 4a that the single-walker runs rarely manage to stretch beyond a distance of $\xi = 22$ Å, whereas the 16-walker simulations explore the whole reaction-coordinate space. Furthermore, in Figure 4b,

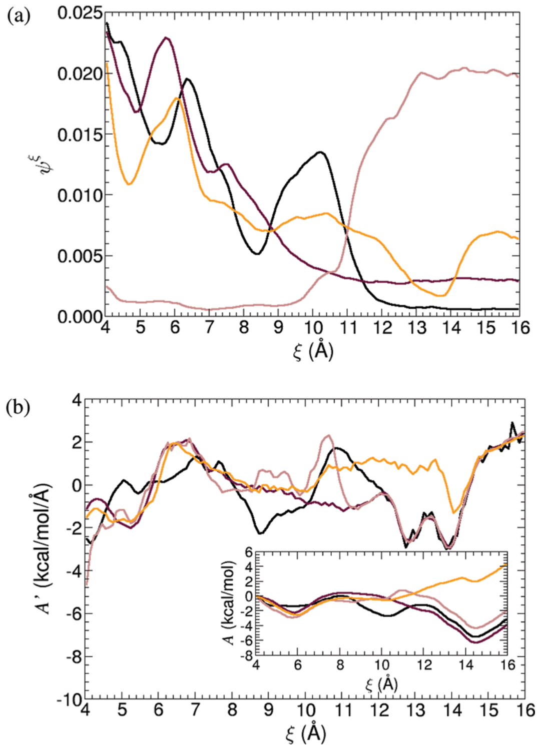


***Figure 5.*** Results for $\xi$ ranging from 4 to 16 Å using 1 walker (after 100 ns). Results are from four independent simulations. (a) Sampling along the RC. (b) Mean force approximations and free-energy profiles (inset): large discrepancies are observed, suggesting the presence of parallel valleys along $\xi$. Note that one of the free-energy profiles suggests a global minimum at $\xi = 6$ Å.

it is apparent that the mean force and free-energy profiles obtained by the 16-walker simulations are already qualitatively consistent with the reference curves.

**5.2. Reaction Coordinate Range: 4−16 Å.** As previously mentioned, convergence of the standard ABF method can be rather slow in the presence of metastabilities on the submanifold of conformations at a fixed value of $\xi$. This is generally the result of a poor choice of the RC, which does not capture all metastabilities of the system. In such a case—as depicted in Figure 1—several low energy conformations could be associated to a fixed value of $\xi$ and separated by high-energy barriers. As highlighted in ref 32, this shortcoming arises when studying compact conformations of the deca-alanine peptide. In this article, an extension of the standard sampling window reveals a free-energy profile that exhibits a wide global minimum ranging from 4 to 12 Å. It is known, however, that the global minimum of the deca-alanine peptide is the α-helical conformation at about $\xi = 14$ Å (see refs 12 and 33). The present results can be explained by the fact that, in compact states, a great number of low-energy conformations are associated to a value of $\xi$ of the RC, which are not fully explored by a standard, single-walker ABF simulation due to their separation by high free-energy barriers. These high free-energy barriers are generally insurmountable from the perspective of conventional MD simulations and can be viewed as kinetic traps that preclude the exploration of the full RC space over reasonable time scales. A recent study has helped to capture the various slow degrees of freedom for these compact structures by exploring multidimensional free-energy landscapes.[13]

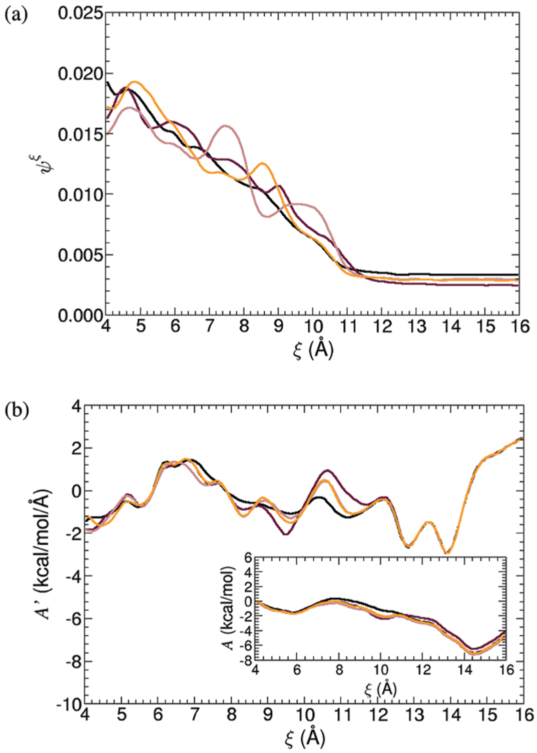


***Figure 6.*** Results for $\xi$ ranging from 4 to 16 Å using 32 walkers (results after 100 ns). Results are from four independent simulations. (a) Sampling along the RC. (b) Mean force approximations and free-energy profiles (inset). Sampling and mean force estimations are consistent with each other, and the α-helical conformation is recovered as the global free-energy minimum.

The shortcomings discussed above can be advantageously circumvented using multiple walkers. The results obtained from 100-ns single- and multiple-walker simulations of the compact conformations are compared in Figures 5 and 6, respectively. Figure 5b depicts mean force estimations for four independent single-walker simulations. Even after a 100-ns simulation, large discrepancies are observed between the mean force profiles. As can be observed in the inset of Figure 5b, one free-energy profile has revealed a global minimum around $\xi = 6$ Å, which is, in most likelihood, artifactual. Figure 6 summarizes the results obtained from four independent 32-walker simulations. A marked improvement in the convergence of the mean force profiles is immediately apparent. This supports the speculation that there exist parallel valleys along the RC, each of a different nature, separated by high free-energy barriers. The present set of results is far more promising with a multiple-walker scheme.

Due to eventual traps in the parallel valleys, it is in fact likely that a $T$-nanosecond single-walker simulation will be less efficient than an $R$-walker simulation run for $T/R$ nanoseconds. This argument is supported numerically by Figure 7, showing results for a 32-walker simulation after 100/32 ∼ 3 ns. The results are qualitatively consistent with Figure 6 and offer a far more reliable set of results than a 100 ns single-walker simulation. In this way, a

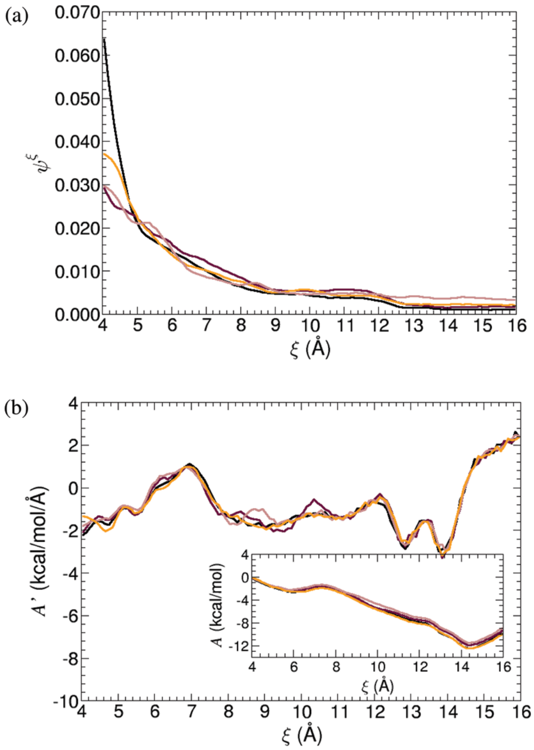


***Figure 7.*** Results after 3 ns using 32 walkers. To compare results at constant total CPU time, we observe the results of a 32-walker simulation after 100/32 ∼ 3 ns. Results are from four independent simulations. (a) Sampling along the RC. (b) Mean force estimates and free-energy profiles (inset) are qualitatively very close to Figure 6b. The results show that a multiple-walker simulation can outperform a single-walker simulation at constant CPU time.

multiple-walker implementation not only improves results at a constant wall time but at a *constant CPU time* as well.

**5.3. Selection.** In order to monitor the impact of the selection mechanism, we have chosen to study again the standard range of 12−32 Å. This choice is dictated by the topology of the free-energy landscape. It ought to be recalled here that the selection criterion is based solely on the position of the walkers along the RC. In the case of multiple, parallel valleys, this criterion can, therefore, impede convergence of the mean force, should many copies of one walker be generated in the same valley. It appears that selection is most effective in the presence of metastabilities in the RC only. Metastabilities in the orthogonal directions are well explored by multiple-walker simulations, albeit not always improved by selection.

Figures 8a and 8b compare the sampling and free-energy profiles determined by a 16-walker simulation after 0.25 ns. We observe a more uniform sampling and a better potential of mean force for the simulation *with* a selection mechanism. For these simulations, we used the selection constant $c = 0.0001$ in eq 23 and a stopping criterion as in eq 25 with $\varepsilon = 0.05$.

Figure 9 depicts $E_w(t)$, the relative entropy of the walker weights, during the first 2.5 ns of the simulation. It may be observed that $E_w(t)$ decreases during the ABF simulation, as the biasing force converges. This result suggests that the walkers are then more free to move along the RC, and thus, each walker is of equal “importance”. Once the walkers are more or less of equal weight, the selection becomes redundant and is, therefore, switched off, avoiding unnecessary computational effort.

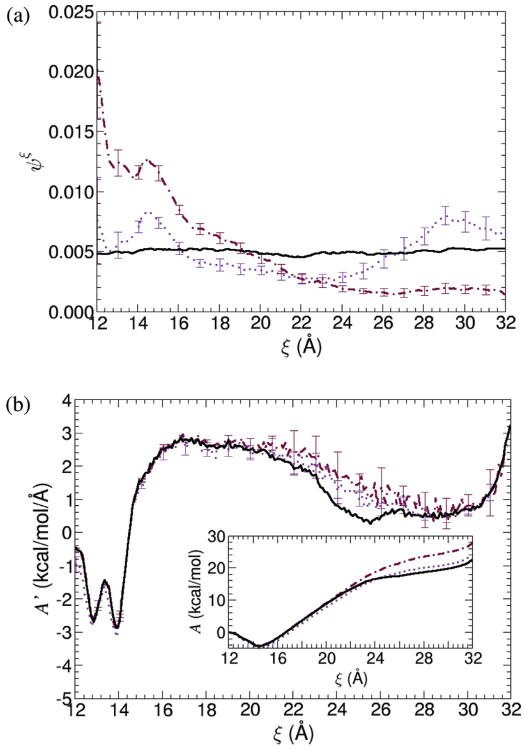


***Figure 8.*** Results for $\xi$ ranging from 12 to 32 Å using 16 walkers (results after 0.25 ns). Comparing results between a 16-walker run with (dotted lines) and without selection (dashed-dotted lines). The curves represent averages of 20 independent ABF simulations, and the error bars are 95% confidence intervals. Reference curves are shown as solid lines. (a) The sampling along $\xi$ shows that simulations *with* selection provide a much more uniform distribution along the RC. (b) Mean force approximations and free energy difference profiles (inset): the free-energy profile for the simulation *with* selection is already very close to the reference curve.

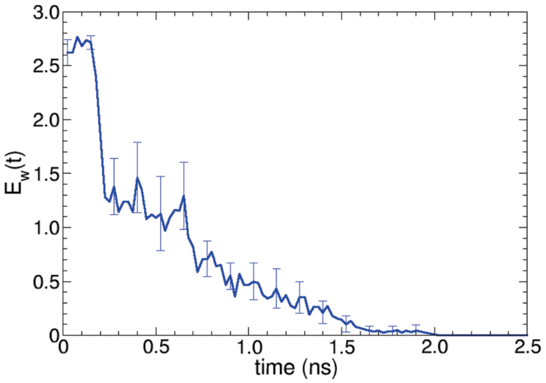


***Figure 9.*** Relative entropy of weights. It can be seen that the $R = 16$ walkers are approximately of equal weight after about 1.5 ns of an ABF simulation. The selection is switched off after $E_w(t) < \varepsilon \log(16)$, where $\varepsilon = 0.05$. For these simulations, the selection constant in eq 23 is chosen as $c = 0.0001$.

## 6. Discussion

In the present contribution, we have demonstrated the applicability of the MW-ABF method to a prototypical biomolecular system. Importance sampling techniques are often held back by the difficulty of choosing good reaction coordinates. If the RC is chosen poorly, one is likely to encounter parallel valleys separated by large free-energy barriers, thereby making sampling at a fixed point along the RC very difficult. In such an event, a standard single-walker ABF simulation would lead to slow convergence, as was shown here. The system is biased only in the direction of the RC and, therefore, would be likely to linger in one valley for a long time before reaching another. We have shown that such shortcomings can be elegantly overcome using multiple walkers, through the proposed MW-ABF method. We emphasize that the use of multiple walkers is particularly beneficial when the choice of the model RC is suboptimal, where improvement has been demonstrated herein at constant CPU time. For a well chosen RC, the MW-ABF is not guaranteed to outperform single-walker ABF simulations at a fixed total CPU cost but still has the advantage of being easily parallelized. The selection process introduced herein can be employed profitably when encountering pronounced free-energy barriers along the RC. In the presence of parallel valleys, attention must be paid to avoid degeneration of weights, as this could lead to many walkers being kinetically trapped in the same valley, losing the main interest of the use of multiple walkers.

**Acknowledgment.** This work is supported in part by the Agence Nationale de la Recherche, under grants LN3M (ANR-05-CIGC-0003) and MEGAS (ANR-09-BLAN-0216-01). The authors are grateful to Gabriel Stoltz and Jérome Hénin for fruitful discussions and to the reviewers for many helpful comments.

### References

(1) Chipot, C.; Pohorille, A. *Free Energy Calculations: Theory and Applications in Chemistry and Biology*; Springer: New York, 2007.

(2) Lelièvre, T.; Rousset, M.; Stoltz, G. *Free Energy Computations: A Mathematical Perspective*; Imperial College Press, to appear.

(3) Zwanzig, R. W. *J. Chem. Phys.* **1954**, *22*, 1420–1426.

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

(5) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011–1021.

(6) Shirts, M. R.; Chodera, J. D. *J. Chem. Phys.* **2008**, *124*, 124105.

(7) Jarzynski, C. *Phys. Rev. Lett.* **1997**, *78*, 2690–2693.

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

(9) Iannuzzi, M.; Laio, A.; Parrinello, M. *Phys. Rev. Lett.* **2003**, *90*, 238302.

(10) Kirkwood, J. G. *J. Chem. Phys.* **1935**, *3*, 300–313.

(11) Darve, E.; Pohorille, A. *J. Chem. Phys.* **2001**, *115*, 9196–9183.

(12) Hénin, J.; Chipot, C. *J. Chem. Phys.* **2004**, *121*, 2904–2914.

(13) Hénin, J.; Fiorin, G.; Chipot, C.; Klein, M. L. *J. Chem. Theory Comput.* **2010**, *6*, 35–47.

(14) Lelièvre, T.; Rousset, M.; Stoltz, G. *J. Comput. Phys.* **2007**, *222*, 624–643.

(15) Bhandarkar, M. *NAMD*; Theoretical Biophysics Group, University of Illinois and Becman Institute: Urbana, IL, 2003.

(16) Kale, L.; Skeel, R.; Bhandarkar, M.; Brunner, R.; Gursoy, A.; Krawetz, N.; Phillips, J. C.; Shinozaki, A.; Varadarajan, K.; Schulten, K. *J. Comp. Phys.* **1999**, *151*, 283–312.

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

(18) Kubo, R.; Toda, M.; Hashitsume, N. *Statistical Physics II: Nonequilibrium Statistical Mechanics*, 2nd ed.; Springer: New York, 1991.

(19) Carter, E. A.; Ciccotti, G.; Hynes, J. T.; Kapral, R. *Chem. Phys. Lett.* **1989**, *156*, 472–477.

(20) Frenkel, D.; Smit, B. *Understanding Molecular Simulation*; Academic Press: New York, 1996.

(21) Ciccotti, G.; Lelièvre, T.; Vanden-Eijnden, E. *Comm. Pure Appl. Math.* **2008**, *61*, 3.

(22) Sprik, M.; Cicotti, G. *J. Chem. Phys.* **1998**, *109*, 7737–7744.

(23) den Otter, W. K.; Briels, W. J. *J. Chem. Phys.* **1998**, *109*, 4139.

(24) Lelièvre, T.; Rousset, M.; Stoltz, G. *Nonlinearity* **2008**, *21*, 1155–1181.

(25) Lelièvre, T.; Rousset, M.; Stoltz, G. *J. Chem. Phys.* **2007**, *126*, 134111.

(26) Assaraf, R.; Caffarel, M.; Khelif, A. *Phys. Rev. E* **2000**, *61*, 4566.

(27) Doucet, A.; de Freitas, N.; Gordon, N. J. *Sequential Monte Carlo Methods in Practice*; Springer: New York, 2001; Series Statistics for Engineering and Information Science.

(28) Del Moral, P. *Feynman-Kac Formulae Genealogical and Interacting Particle Systems with Applications*; Springer: New York, 2004.

(29) Douc, R.; Cappe, O.; Moulines, E. *Image and Signal Processing and Analysis, 2005. ISPA 2005* **2005**, 64–69.

(30) Kitagawa, G. *J. Comput. Graph. Stat.* **1996**, *5*, 1–25.

(31) Carpenter, J.; Clifford, P.; Fearnhead, P. *Technical Report, Department of Statistics*; University of Oxford: Oxford, U.K., 1999.

(32) Chipot, C.; Hénin, J. *J. Chem. Phys.* **2005**, *123*, 244906.

(33) Park, S.; Khalili-Araghi, F.; Tajkhorshid, E.; Schulten, K. *J. Chem. Phys.* **2003**, *119*, 3559.

(34) MacKerell, A. D.; et al. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

CT900524T

JCTC Journal of Chemical Theory and Computation

# Effect of the Integration Method on the Accuracy and Computational Efficiency of Free Energy Calculations Using Thermodynamic Integration

Miguel Jorge,*,† Nuno M. Garrido,† António J. Queimada,† Ioannis G. Economou,‡,§ and Eugénia A. Macedo†

*LSRE Laboratory of Separation and Reaction Engineering, Departamento de Engenharia Química, Faculdade de Engenharia, Universidade do Porto, Rua do Dr. Roberto Frias, 4200 - 465 Porto, Portugal, Molecular Thermodynamics and Modeling of Materials Laboratory, Institute of Physical Chemistry, National Center for Scientific Research "Demokritos", GR-15310 Aghia Paraskevi Attikis, Greece, The Petroleum Institute, Department of Chemical Engineering, P.O. Box 2533, Abu Dhabi, United Arab Emirates*

Received December 9, 2009

**Abstract:** Although calculations of free energy using molecular dynamics simulations have gained significant importance in the chemical and biochemical fields, they still remain quite computationally intensive. Furthermore, when using thermodynamic integration, numerical evaluation of the integral of the Hamiltonian with respect to the coupling parameter may introduce unwanted errors in the free energy. In this paper, we compare the performance of two numerical integration techniques—the trapezoidal and Simpson's rules—and propose a new method, based on the analytic integration of physically based fitting functions that are able to accurately describe the behavior of the data. We develop and test our methodology by performing detailed studies on two prototype systems, hydrated methane and hydrated methanol, and treat Lennard-Jones and electrostatic contributions separately. We conclude that the widely used trapezoidal rule may introduce systematic errors in the calculation, but these errors are reduced if Simpson's rule is employed, at least for the electrostatic component. Furthermore, by fitting thermodynamic integration data, we are able to obtain precise free energy estimates using significantly fewer data points (5 intermediate states for the electrostatic component and 11 for the Lennard-Jones term), thus significantly decreasing the associated computational cost. Our method and improved protocol were successfully validated by computing the free energy of more complex systems—hydration of 2-methylbutanol and of 4-nitrophenol—thus paving the way for widespread use in solvation free energy calculations of drug molecules.

## 1. Introduction

Calculation of free energies is extremely important for a wide spectrum of technological areas, perhaps most notably in the pharmaceutical industry, where solvation free energy estimates are essential to predict, for example, drug solubility and protein−ligand binding energies.[1,2] Thus, computational methods that are able to predict accurate solvation free energy values can bring tremendous advances in drug design methodologies. With recent improvements in computer power and algorithms, molecular simulation-based free energy calculations are being performed in a more routine way (as an example, Mobley et al. recently calculated the hydration free energy of 504 compounds using molecular

* Author to whom all correspondence should be addressed. E-mail: mjorge@fe.up.pt.

† Universidade do Porto.

‡ National Center for Scientific Research "Demokritos".

§ The Petroleum Institute.

10.1021/ct900661c © 2010 American Chemical Society
Published on Web 02/24/2010

simulation[3]). Nevertheless, we have not yet reached a stage where these methods are predictive enough for practical use.[4] A major stumbling block is the fact that the parametrization of most molecular force fields does not take free energy data into account (a notable exception being the recent parametrizations of the GROMOS force field[5]), which is understandable given that such calculations are still much more computationally demanding than calculations of bulk fluid properties and phase equilibria. There is thus a pressing need to make free energy calculation methods as fast as possible. Furthermore, such calculations must be very precise—if the error intrinsic to the calculation method is small (high precision), any differences between simulation and experiment can be confidently attributed to inaccuracies in the molecular model, which can then be appropriately refined. The problem is that precision and speed do not normally come hand-in-hand, and in practice one must find an appropriate balance between the two. In this work, we explore different integration methods in an attempt to improve both the precision and the speed of free energy calculations using Thermodynamic Integration (TI) of molecular simulation data.

TI, originally proposed by Kirkwood,[6] is the most widely used, and perhaps most robust, method for computing solvation free energies of complex solutes (for a review of other methods and a more detailed description of TI, the reader is referred, for example, to the recent book by Chipot and Pohorille[7]). The TI method considers a transition between two generic well-defined states, an initial reference state (state 0) and a final target state (state 1), described by the Hamiltonians $H_0$ and $H_1$, respectively. A coupling parameter, $\lambda$, is added to the Hamiltonian, $H(\boldsymbol{p},\boldsymbol{q};\lambda)$, where $\boldsymbol{p}$ is the linear momentum and $\boldsymbol{q}$ the atomic position, and used to describe the transition between the end-points: $H(\boldsymbol{p},\boldsymbol{q};0)\rightarrow H(\boldsymbol{p},\boldsymbol{q};1)$. Considering several discrete and independent $\lambda$ values between 0 and 1, equilibrium averages can be used to evaluate derivatives of the free energy with respect to $\lambda$. One then integrates the derivatives of the free energy along a continuous path connecting the initial and final states in order to obtain the energy difference between them:

$$\Delta G = \int_0^1 \left\langle \frac{\partial H(\boldsymbol{p},\boldsymbol{q},\lambda)}{\partial \lambda} \right\rangle_{\lambda} \mathrm{d}\lambda \qquad (1)$$

where the angular brackets indicate an ensemble average at a particular value of $\lambda$. Equation 1 is exact but suffers from two possible sources of error: (i) the statistical error in the ensemble average of the Hamiltonian derivative at each value of $\lambda$ and (ii) the error associated with the integration of the curve. The first error can be reduced, in principle, by increasing the length of each individual simulation. The second type of error is normally addressed by increasing the number of intermediate points. Indeed, it has been concluded that the precision of the TI methodology depends mostly on the smoothness of the $\partial H/\partial\lambda$ vs $\lambda$ plot.[8] As a rule of thumb, it was suggested that the free energy difference between two consecutive points ($\lambda$ and $\lambda + \Delta\lambda$) should be less than 2 kcal/mol.[9] If we deal with a system containing high energy barriers, the number of intermediate steps may become considerably large and the associated computational cost too high. Here, we analyze in detail the impact of the choice of integration method and the number of intermediate points on the precision of the free energy estimate.

The trapezoidal rule is by far the most widely used method to numerically evaluate the integral in eq 1 when estimating $\Delta G$ via TI. A notable exception is the use of Gauss–Legendre integration in the work of Smith et al.[10] The trapezoidal rule performs a linear interpolation between successive points and can thus suffer from systematic errors if the underlying function is very far from linearity (which is indeed the case for most practical calculations). An alternative to reduce such deviations is to use a more elaborate integration method, such as the Simpson rule. However, to our knowledge, this has not been previously explored in free energy calculations. Another option would be to fit the entire data set to an appropriate functional form and then perform the integration of this function analytically. This idea has been applied before by Swope and Andersen[11] where average solute–water interactions in the hydration of inert gases were fitted as a function of the coupling parameter and by Hummer and coworkers in the context of charging free energies.[12] Recently, while this manuscript was being prepared, Shyu and Ytreberg[13] demonstrated that the use of polynomial functions to fit simulation data can significantly increase the precision of the free energy estimates over the trapezoidal rule, without requiring additional simulations. However, they have examined only very simple prototype systems, with an analytical solution to the free energy and smooth monotonous curves. As we will show below, simple polynomial functions are not the best choice to describe the curves that arise in hydration free energy calculations, even for small solutes.

In the present work, we compare the performance of two numerical integration techniques—the trapezoidal rule and Simpson's rule—in the calculation of free energies from TI. Furthermore, we develop a physically based fitting function that is able to accurately describe the variation of the Hamiltonian derivative with respect to the coupling parameter. By fitting this function to the simulation data, we are able to obtain precise free energies using significantly fewer intermediate points, thus decreasing the associated computational cost. We carry out our detailed study for two prototype systems, methane and methanol in water, which represent realistic solutes (both polar and apolar) and a realistic solvent, but are simple enough to allow for long simulations to be performed at a very large number of intermediate values of $\lambda$, an essential requisite to assessing the validity of our procedure. We then apply our methodology to the solvation of two larger and more complex molecules, namely, 2-methylbutanol and 4-nitrophenol, in order to demonstrate its applicability in realistic free energy calculations. In the following section, we present a detailed description of the simulation methods, while the integration methods and the development of the fitting function are explained in section 3. Section 4 presents the results of our study followed by the main conclusions in section 5.

## 2. Computational Details

Molecular dynamics (MD) simulations were performed using the GROMACS simulation suite.[14] Hydrated systems consisted of one solute molecule (methane, methanol, 2-methylbutanol, or 4-nitrophenol) represented by the OPLS-AA[15] force field and 500 water molecules represented by the SPC/E[16] model (parameters for the models are provided in Tables S1−S4 of the Supporting Information). Covalent bonds involving hydrogen atoms were constrained with the LINCS[17] algorithm, while the water geometry was fixed with the SETTLE[18] algorithm. For efficiency reasons,[19] we have used the reaction-field method, originally proposed by Lee et al.,[20] with a cutoff distance of 1 nm and a dielectric permittivity of 80, to account for long-range electrostatic interactions. The remaining cutoff radii used were 1 nm for the short-range neighbor list and a 0.8−0.9 nm switched cutoff for the Lennard-Jones (LJ) interactions. We have applied long-range corrections for energy and pressure as suggested in the work of Shirts et al.[8] Simulations were performed using periodic boundary conditions in all directions. Newton's equations of motion for all species were integrated using the leapfrog dynamic algorithm[21] with a time step of 2 fs. Langevin stochastic dynamics[22] was used to control the temperature, with a frictional constant of 1 ps$^{-1}$, while for constant pressure runs, the Berendsen barostat,[23] with a time constant of 0.5 ps and an isothermal compressibility of $4.5 \times 10^{-5}$ bar$^{-1}$, was used to enforce pressure coupling.

The TI method makes use of a thermodynamic cycle to compute the free energy required to transfer a given solute from the gas phase to the solvent. The three stages of the cycle are (i) transforming the solute into a dummy molecule (i.e, turning off all nonbonded interactions) in a vacuum, (ii) solvating the dummy molecule, and (iii) transforming the dummy molecule into the solute in water. Because dummy molecules have no interactions with their environment, the free energy associated with stage ii is zero by definition. Stage i is normally required to compensate for intramolecular interactions that are coupled to the nonbonded parameters. However, methane is small enough that this contribution is zero (there are no atoms separated by more than two bonds). In the other three solutes, vacuum calculations need to be performed because 1−4 interactions are present, but for the LJ component of methanol, these turn out to be zero as well (the LJ parameters for the hydroxyl hydrogen atom are zero in the OPLS-AA model[15]). For stages i and iii, the total solvation free energy can be calculated by transforming the fully interacting solute ($\lambda = 1$) into a dummy solute ($\lambda = 0$) in a vacuum and in water, respectively. In the case of methanol, 2-methylbutanol and 4-nitrophenol (polar solutes), this operation was performed in two steps—first the charges were gradually turned off and then the LJ parameters were decoupled—thus avoiding charge fusion effects.[8] A linear dependence of the electrostatic interactions with the coupling parameter was imposed. For all four solutes, the soft-core function of Beuler et al.[24] was used for the dependence of the LJ term with $\lambda$:

$$V_{\mathrm{SC}} = \lambda V[(\alpha\sigma^6(1 - \lambda)^p + r^6)^{1/6}] \tag{2}$$

In this equation, $V(r)$ is the normal "hard-core" pair potential, $\alpha$ is the soft-core parameter, and $\sigma$ is the LJ site diameter. This soft-core dependence eliminates singularities in the calculation as the LJ interactions are turned off and is the only scaling protocol that yields completely stable dynamics near the end points, as reported in a comparison of different nonbonded scaling approaches for free energy calculations.[25] We have used a value of $p = 1$ for the power of the $\lambda$ dependence, since this produces a much smoother $\partial H/\partial\lambda$ for LJ interactions.[8] The value of $\alpha$ was 0.5, which is the optimized value for $p = 1$, as reported by Mobley et al.[26]

Initial configurations for each point were generated by immersing the solute molecules in a previously equilibrated water box at 298 K and 1 bar, after which short equilibration runs were performed. For each simulation, we then ran an energy minimization (using the limited-memory Broyden−Fletcher−Goldfarb−Shanno algorithm[27] over 5000 steps followed by a steepest descent minimization of 1000 steps) followed by a constant volume equilibration (100 ps), a constant pressure equilibration (500 ps) long enough to obtain complete equilibration of the box volume, and finally a 5 ns *NpT* production stage. This procedure was repeated for each $\lambda$ value, allowing for a separate minimization. Sampling errors for each individual simulation were estimated using the block averaging procedure of Flyvbjerg and Petersen.[28] For the purpose of our study, it is important to have a very precise estimate of $\Delta G$ to serve as a reference value. To achieve this, we have used a total of 129 equidistant intermediate points for each of the small solutes (for both LJ and electrostatic components in the case of methanol). Equidistant points are preferable when there is no *a priori* knowledge of the final shape of the $\partial H/\partial\lambda$ plot. Reduced data sets were built by manipulating the original set of 129 points, as described in section 4. For 2-methylbutanol and 4-nitrophenol, we used 31 points for the LJ component and 17 points for the electrostatic component, as explained in detail below.

## 3. Integration Methods

The simplest method to integrate a curve composed of discrete points is the trapezoidal rule. This is a first order method, which simply interpolates linearly between consecutive values of $x$, resulting in the following generic formula:

$$\int_{x_1}^{x_N} f(x)\,\mathrm{d}x = \sum_{i=1}^{N-1} (x_{i+1} - x_i)\frac{f(x_{i+1}) + f(x_i)}{2} \tag{3}$$

where $N$ is the total number of points in the required interval and $f(x)$ is the function one wishes to integrate. The trapezoidal rule can be applied with any number of points separated by any distance. In the special case of evenly distributed points in the integration interval, eq 3 simplifies to

$$\int_{x_1}^{x_N} f(x)\,\mathrm{d}x = h\left[\frac{f(x_1)}{2} + \sum_{i=2}^{N-1} f(x_i) + \frac{f(x_N)}{2}\right] \tag{4}$$

where $h$ is the interval between two consecutive points. Due to its simplicity and versatility, the trapezoidal rule is widely employed and has been the method of choice in the large majority of free energy calculations by thermodynamic integration.

A more accurate integration method is Simpson's rule. It is a second order method (i.e., interpolates between 3 successive points using a quadratic polynomial) but turns out to be exact up to degree 3 due to a cancellation of coefficients.[29] The generic formula is

$$\int_{x_1}^{x_N} f(x)\,\mathrm{d}x = \sum_{i=1}^{(N-1)/2} (x_{2i+1} - x_{2i-1})\frac{f(x_{2i-1}) + 4f(x_{2i}) + f(x_{2i+1})}{3} \qquad (5)$$

Notice that Simpson's rule requires that $N$ be odd (i.e., an even number of intervals) and that any three successive points be separated by equal intervals. In practice, however, it is almost always applied to situations in which the points are all evenly distributed in the integration interval. In this case, eq 5 reduces to

$$\int_{x_1}^{x_N} f(x)\,\mathrm{d}x = \frac{2h}{3}[f(x_1) + 4f(x_2) + \sum_{i=3}^{N-1}[3 + (-1)^i]f(x_i) + f(x_N)] \qquad (6)$$

An alternative to the above numerical integration schemes is to use a fitting function. In this case, a specific functional form, with a certain number of fitting parameters (as few as possible), is fitted through all the data points in the integration interval, and the desired integral is then evaluated directly from the fitting function. The simplest fitting functions that can be applied are polynomials of the form

$$\int_{x_1}^{x_n} f(x)\,\mathrm{d}x = \sum_{i=1}^{n_\mathrm{P}} a_i x^i \qquad (7)$$

where $n_\mathrm{P}$ is the degree of the fitting polynomial and $a_i$ are the unknown coefficients (i.e., the fitting parameters). Notice that the term for $i = 0$ is taken to be zero, so that the function passes through the origin. Normally, increasing $n_\mathrm{P}$ leads to a better fit of the data set that one wishes to integrate. In practice, however, a point is usually reached when the error of the polynomial expansion is on the same order as the uncertainty in the data, and a further increase of $n_\mathrm{P}$ leads to no improvement of the fit. Notice also that, in order for the fitting to be meaningful, one must always have $N \geq n_\mathrm{P}$. A further problem with polynomial fits is that they tend to produce unphysical oscillations for data sets that show a complicated dependence on $x$.[29]

As we will see below, polynomial functions provide an excellent description of the electrostatic contribution to the free energy but are inappropriate for fitting the Lennard-Jones component, due to the more complicated dependence on $\lambda$. In the latter case, we have searched for a more physically based fitting function. The reader is warned that the following is not meant to be a rigorous model for describing the LJ contribution to the hydration free energy but is simply a method of obtaining a fitting function that is based on the physics of that contribution. Indeed, it involves some very crude assumptions regarding the nature of the interactions in the system but is nevertheless able to yield a good fit of the LJ data, as we will see below.

The total LJ contribution to the free energy may be considered to arise from a competition between two different components, one due to (unfavorable) cavity formation in the solvent and the other due to (favorable) van der Waals interactions between the solute and solvent.[30] The first component is mainly entropic in nature and is predominant at small values of $\lambda$, while the second component is mainly enthalpic and dominates for large values of $\lambda$. The cavity formation free energy may be expressed as the sum of a volume term (the work acting against an external pressure) and a surface term (work acting against the surface tension), as follows:[30,31]

$$\Delta G_\mathrm{Cav} \sim \frac{4\pi}{3} p r^3 \lambda^3 + 4\pi\gamma r^2 \lambda^2\left(1 - \frac{4\delta}{r\lambda}\right) \qquad (8)$$

where $p$ is the pressure, $r$ is the solute radius, $\gamma$ is the surface tension, and $\delta$ is a curvature correction to the surface tension. A similar expression can be derived from scaled-particle theory:[30,32]

$$\Delta G_\mathrm{Cav} \sim K_3\lambda^3 + K_2\lambda^2 + K_1\lambda + K_0 \qquad (9)$$

Taking any of these forms, it is easy to see that the cavity contribution to the Hamiltonian derivative can be approximated by a quadratic expression:

$$\left(\frac{\partial H}{\partial \lambda}\right)_\mathrm{Cav} = A_0\lambda^2 + A_1\lambda + K \qquad (10)$$

where we take $A_0$, $A_1$, and $K$ as adjustable (free) parameters.

As for the attractive term, it is reasonable to assume that, once the cavity is formed, there will be no significant solvent restructuring caused by turning on the attractive interactions.[30,33] This mean-field approximation implies that the entropic contribution is negligible, and thus the free energy is given simply by the solute−solvent van der Waals interaction energy. Furthermore, we introduce the simplification that this attractive energy is the sum of an explicit and an implicit term, as follows:

$$\left(\frac{\partial H}{\partial \lambda}\right)_\mathrm{Attr} \sim \frac{\partial E_\mathrm{LJ}}{\partial \lambda} = \frac{\partial E_\mathrm{Expl}}{\partial \lambda} + \frac{\partial E_\mathrm{Impl}}{\partial \lambda} \qquad (11)$$

The explicit term contains the contributions from the first solvation shell of water molecules around the solute, while the implicit term contains the contributions of all other water molecules in the system. We approximate the implicit term by a continuum, obtained by integrating the attractive part of the LJ potential between a distance $R_\mathrm{C}$ and infinity:

$$E_\mathrm{Impl} = \int_{R_C}^{\infty} 4\pi r^2 V_\mathrm{LJ}(r)\,\mathrm{d}r \qquad (12)$$

Substituting the attractive part of the LJ potential in the above equation and integrating, we obtain

$$E_\mathrm{Impl} = -\int_{R_\mathrm{C}}^{\infty} 16\pi\varepsilon\lambda\sigma^6 r^{-4}\,\mathrm{d}r = -\frac{16\pi\varepsilon\sigma^6}{3R_\mathrm{C}^3}\lambda \qquad (13)$$

where $\sigma$ and $\varepsilon$ are the LJ solute−solvent diameter and well depth, respectively. The derivative of eq 13 with respect to $\lambda$ yields a constant term, as expected.

Regarding the explicit term, we make the rather crude assumption that all the $n_\mathrm{W}$ water molecules in the first solvation shell are at the same distance $R$ from the solute. With this assumption, the potential energy is given simply by the attractive term multiplied by $n_W$. Here, we must take the soft-core expression, eq 2, for the attractive term:

$$E_{\mathrm{Expl}} = -\frac{4\varepsilon\sigma^6 n_W \lambda}{\alpha\sigma^6(1-\lambda)^p + R^6} \tag{14}$$

Taking the derivative with respect to $\lambda$ yields

$$\frac{\partial E_{\mathrm{Expl}}}{\partial\lambda} = -4\varepsilon n_W \frac{\alpha p\lambda(1-\lambda)^{p-1} + \alpha(1-\lambda)^p + (^R/_\sigma)^6}{[\alpha(1-\lambda)^p + (^R/_\sigma)^6]^2} \tag{15}$$

By taking $p = 1$ for the soft-core power (see section 2) and expanding, we obtain an expression of the form:

$$\left(\frac{\partial H}{\partial\lambda}\right)_{\mathrm{Attr}} = \frac{-A_2}{\lambda^2 - A_3\lambda + A_4} - B \tag{16}$$

where once more we take $A_2$, $A_3$, $A_4$, and $B$ as adjustable parameters. Now all we need to do is combine eqs 10 and 16 to obtain a fitting function for the Hamiltonian derivative. Before we do that, however, we introduce an additional requirement:

$$\lim_{\lambda\to 0}\left(\frac{\partial H}{\partial\lambda}\right) = 0 \Rightarrow K - B = \frac{A_2}{A_4} \tag{17}$$

This means that all the constant terms will cancel out and the curve will go through zero at $\lambda = 0$. The final expression, with 5 adjustable parameters, is

$$\left(\frac{\partial H}{\partial\lambda}\right)_{\mathrm{LJ}} = A_0\lambda^2 + A_1\lambda - \frac{A_2}{\lambda^2 - A_3\lambda + A_4} + \frac{A_2}{A_4} \tag{18}$$

Equation 18 has an analytic integral that depends on the nature of the roots of the quadratic expression in the denominator of the third term. In fact, if any of the roots falls between 0 and 1, the function will have a discontinuity in our region of interest. To avoid this, we can require that the discriminant of the polynomial always be negative, so that both roots are complex. This means adding the following constraint to the fitting procedure:

$$U = 4A_4 - A_3^2 > 0 \tag{19}$$

In practice, we found out that a strict use of this (unnecessarily strong) constraint was not needed, provided that the initial estimate of parameters $A_3$ and $A_4$ obeyed the above inequality. When eq 19 is obeyed, the integral of eq 18 between 0 and 1 is given by

$$\Delta G_{\mathrm{LJ}} = \frac{A_2}{A_4} + \frac{A_0}{3} + \frac{A_1}{2} + \frac{2A_2}{\sqrt{U}}\left[\arctan\left(-\frac{A_3}{\sqrt{U}}\right) + \arctan\left(\frac{(A_3 - 2)}{\sqrt{U}}\right)\right] \tag{20}$$

All fits were performed using a nonlinear weighted least-squares routine, as implemented in the xmGrace software.[34]

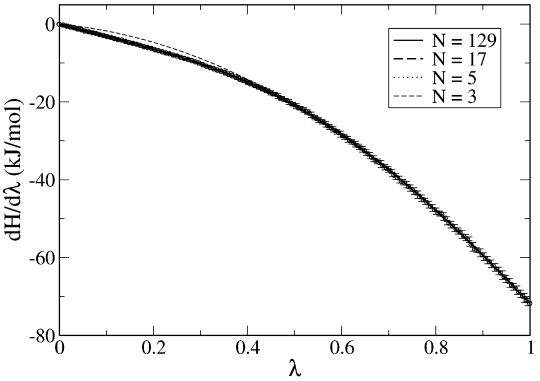


***Figure 1.*** Electrostatic contribution (vacuum − water) to the derivative of the Hamiltonian with respect to $\lambda$ for methanol (open circles with error bars). The lines are fits to the full and reduced data sets using a quartic polynomial function.

***Table 1.*** Results Obtained by Fitting the Electrostatic Contribution Data for Methanol to Polynomial Functions of Increasing Degree ($n_P$)

| $n_P$ | rms error | $\chi^2/(N - n_P)$ | $\Delta G_{\mathrm{Elec}}$ (kJ/mol) | $\epsilon_R$ (%) |
|---|---|---|---|---|
| 2 | 0.1892 | 104.6 | −26.225 | 0.690 |
| 3 | 0.1132 | 16.59 | −26.351 | 0.216 |
| 4 | 0.0899 | 0.997 | −26.407 | 0.002 |
| 5 | 0.0895 | 0.872 | −26.406 | 0.007 |
| 6 | 0.0894 | 0.516 | −26.401 | 0.026 |

## 4. Results and Discussion

**4.1. Electrostatic Component.** We begin by analyzing the electrostatic contribution to the hydration energy of methanol (for the nonpolar methane molecule, this contribution is zero). The data for the total contribution (i.e., vacuum − water) are presented in Figure 1 for the 129 $\lambda$ values considered. The full data set together with the corresponding standard deviations for each simulation are given in the Supporting Information, Table S5. As we can see, the curve is smooth and monotonic, and the sampling error is rather small for all data points. Linear response theory predicts a quadratic dependence of the free energy with respect to the solute charge,[12] which results in a linear dependence for the derivative of the free energy with respect to $\lambda$. However, the data of Figure 1 exhibit significant deviations from linearity and thus suggest a breakdown in linear response theory. This may be attributed to the fact that the solvent is not a uniform dielectric, and thus specific interactions between the solute and the solvent invalidate the linear coupling assumption. This was also verified in other works, e.g., for the charging/uncharging of simple molecules, such as monatomic ions,[9] or for more complex molecules.[8] Indeed, our data could not be accurately fitted using either a linear or a quadratic expression, even for a solute as simple as methanol, and the departure from linear behavior is expected to increase as the solute becomes more complex.

We have fitted the data of Figure 1 to polynomials of increasing degree, following eq 7, and the results are shown in Table 1 (the respective fits are depicted in the Supporting Information, Figure S1). It is clear that the root-mean-square (rms) error of the fit decreases significantly from a quadratic to a quartic polynomial but then shows no significant change

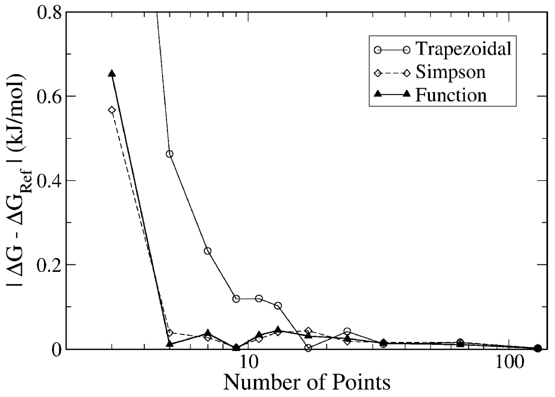


***Figure 2.*** Absolute error in the electrostatic contribution to the free energy, relative to the result for the full data set, as a function of the number of points used in the integration. Open circles are for the trapezoidal rule, open diamonds for Simpson's rule, and full triangles for the analytical integration of the fitting function.

as $n_P$ is further increased. A statistical estimate of the quality of the fit is given by the $\chi^2$ value, which should be on the same order as the number of degrees of freedom of the fit[29] (in this case, $N-n_P$). The improvement is remarkable upon increasing $n_P$ from 2 to 4, but there is only a small change by further increasing the polynomial degree. Finally, the error in the value of the integral computed analytically from the fitting function relative to the value calculated by numerical integration of the data using Simpson's rule, denoted as $\epsilon_R$, actually shows a minimum at $n_P = 4$. This analysis leads us to conclude that the electrostatic contribution curve is ideally fitted by a polynomial of degree 4.

Now that we have established the optimal fitting function, it is time to compare the precision of the numerical methods with the analytic integration as the number of data points ($N$) is reduced. For this purpose, we have generated reduced data sets with fewer $\lambda$ values by removing points from the full data set, such that the points in the reduced data sets were spaced as evenly as possible. Most of these reduced sets (i.e., with $N = 65$, 33, 17, 9, 5, and 3) were generated by dividing the original number of intervals (128) by successive powers of 2, and so the points were all evenly distributed. For the other reduced sets (i.e., $N = 24$, 13, 11, and 7), only one or two points at the extremities of the integration range were not evenly distributed. For each of the reduced data sets, the free energy was computed both numerically, using either the trapezoidal rule, eq 4, or Simpson's rule, eq 6, and analytically, after fitting the data set to a quartic polynomial. The fits using some of the reduced data sets, as well as for the full set, are shown as lines in Figure 1. In Figure 2, we plot the absolute error in the free energy, relative to the reference case (numerical integration with Simpson's rule using the complete 129-point data set), as a function of the number of points in the data set, for the three integration methods considered. The full results of our analysis of the electrostatic component, including values of the fitting parameters, $\chi^2$ values for the fits, and total free energies, are given in Supporting Information, Table S8.

Analyzing Figure 1, we can see that with as few as 5 evenly spaced data points, the behavior of the entire curve is well captured by the fitting function. When $N$ is reduced even further, one runs into overfitting problems, i.e., the polynomial degree is higher than the number of data points available for the regression. In this situation, the number of degrees of freedom of the fit exceeds the information content of the data, and there is arbitrariness in the final fitting model. Indeed, for the data set with 3 points, we have used a quadratic function, rather than a quartic—as can be seen from Figure 1, the results are not very satisfactory.

From Figure 2, we can see that using up to 17 points all three methods yield free energies that are within 0.05 kJ/mol from the reference value. However, if the number of points is reduced further, the error of the trapezoidal rule increases significantly. Remarkably, both the Simpson rule and the analytic integral based on the fitting function perform extremely well down to 5 data points. This is understandable if we consider the shape of the curve (Figure 1)—the convex shape and monotonic behavior means that the linear interpolation between successive points that is at the core of the trapezoidal rule will produce a systematic underestimation of the free energy. Naturally, this systematic error can be reduced by increasing the number of points. On the contrary, both the piecewise quadratic interpolation of Simpson's rule and the quartic polynomial fit are able to correctly capture the curvature of the data and require only a very small number of intermediate points to yield a precise free energy value. This finding is quite important if we take into account that the large majority of calculations of the electrostatic contribution to the free energy are carried out with fewer than 17 points and using the trapezoidal rule to compute the integral. Thus, it is likely that most results in the literature present a systematic bias that may be quite significant.

**4.2. Lennard-Jones Component.** We turn now to an analysis of the integration of the LJ contribution to the free energy. The full data sets, including the corresponding standard deviations, are provided in Table S5 (Supporting Information) and plotted in Figure 3 for both methane and methanol. The curve for the LJ contribution is dominated by a prominent peak located between 0.2 and 0.3 for both solutes; it first increases smoothly at low values of $\lambda$ and decreases again smoothly after the peak. This shape is much more complex than for the electrostatic contribution (Figure 1). It is also important to notice that the sampling errors are also much larger than for the electrostatic contribution, particularly in the vicinity of the peak. This is shown more clearly in Figure S2 of the Supporting Information. The behavior of the LJ curve reflects two competing factors: unfavorable excluded volume effects due to cavity creation in the solvent and favorable solute−solvent interactions.[35] This interpretation has formed the basis for our development of the fitting function, eq 18. In fact, it is important to notice that the data to the left of the peak are very well fitted by our partial expression for the cavity formation term, eq 10, while the data to the right of the peak are well described by the expression derived for the attractive term, eq 16. These partial fits to the data, depicted in Figure 4 for the case of

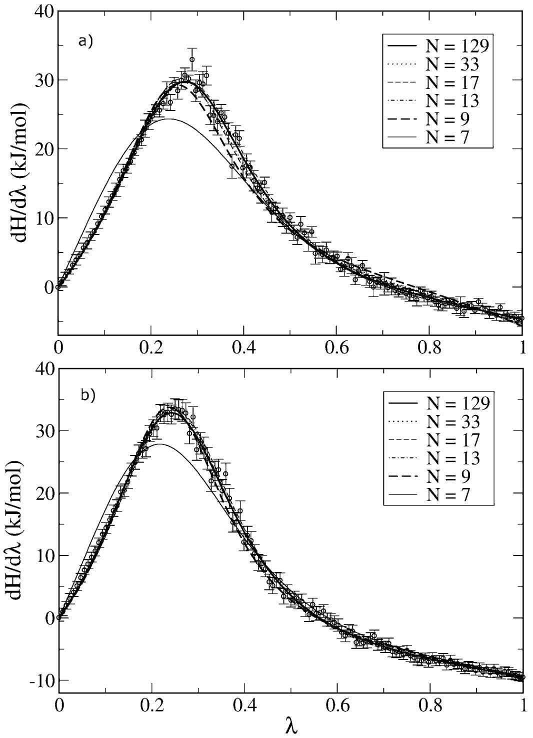


***Figure 3.*** Lennard-Jones contribution to the Hamiltonian derivative with respect to $\lambda$ for (a) methane and (b) methanol (open circles with error bars). The lines are fits to the full and reduced data sets, as indicated, using eq 18.

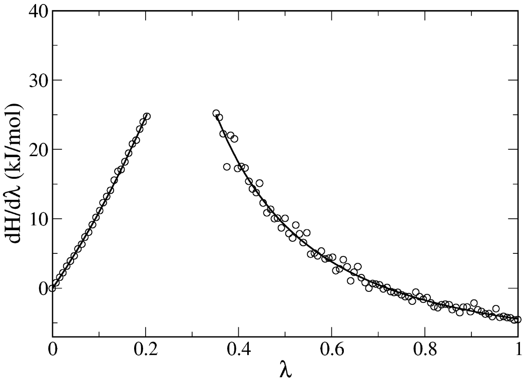


***Figure 4.*** Partial fits to the LJ contribution for methane. The data at low $\lambda$ were fitted to eq 10, while the data for high $\lambda$ were fitted to eq 16.

methane, validate our approach in developing the fitting function for the LJ contribution to the free energy.

The full data sets were fitted to eq 18, and the results are shown as thick lines in Figure 3. As we can see, the function is able to correctly describe the data in the entire region of interest, despite the large amount of statistical noise in the vicinity of the peak. Using the same procedure as in the case of the electrostatic component, we have generated reduced data sets and carried out the integration using the two numerical methods and the fitting function. The fitted curves are shown as lines in Figure 3, while the full results of the analysis, including values of the fitting parameters, $\chi^2$ values for the fits, and total free energies, are provided in the Supporting Information, Tables S9 and S10. In Figure 5, we show the absolute error in the free energy, relative to the reference case, as a function of the number of points in the data set, for the three integration methods considered.

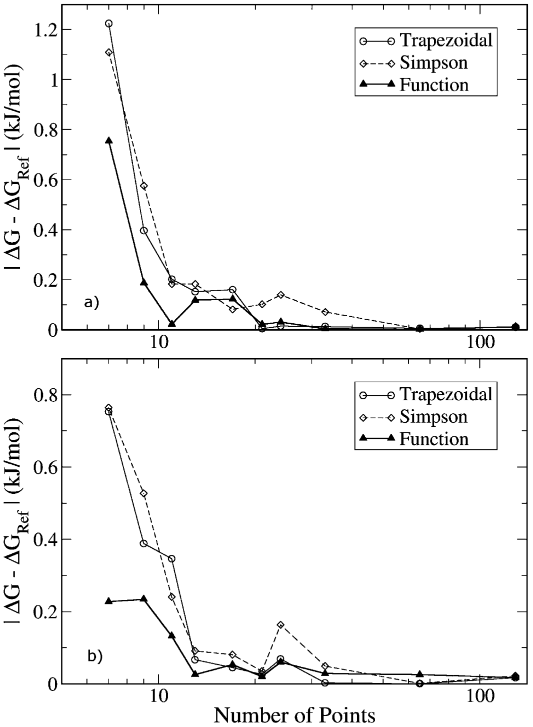


***Figure 5.*** Absolute error in the LJ contribution to the free energy, relative to the result for the full data set, as a function of the number of points used in the integration, for (a) methane and (b) methanol. Open circles are for the trapezoidal rule, open diamonds for Simpson's rule, and full triangles for the analytical integration of the fitting function.

It is clear from Figure 3 that the fitting function is able to correctly describe the trend of the Hamiltonian derivative even using only a small number of points in the fit (a good description is obtained with as few as 11 points). With 9 points, the fitted curve starts to deviate significantly from the full data set, particularly in the case of methane (see thick dashed line in Figure 3a), and with 7 points the performance is quite poor. The performance of the different integration methods can be assessed quantitatively by analyzing Figure 5. First of all, it is worth noticing that in general the errors are larger and show more scatter than for the electrostatic component, which is caused by the higher degree of statistical noise in the simulated data. Furthermore, Simpson's rule now does not significantly outperform the trapezoidal rule—since the function has a maximum, the systematic error of the trapezoidal rule tends to cancel out after the full integration. As expected, the error tends to increase as the number of points is reduced, but this increase is not very pronounced down to $N = 17$. In this region, all three integration methods show a similar performance. As the number of points is reduced further, the error of both numerical integration schemes increases significantly. Using the fitting function, however, one is able to maintain a good precision down to about 11 points, and the difference relative to the numerical methods is even more marked for 9 points. Probably the most important conclusion of our analysis is that when considering a small number of intermediate stages (we recommend using 11 for the LJ contribution) the fitting function always produces more precise results than the two numerical integration techniques.

At this point, it is worth commenting on the possibility of using different fitting functions for the LJ component. Shyu and Ytreberg[13] have performed a systematic analysis of polynomial fits to free energy data but have only applied their procedure to simple test cases with monotonous curves and analytical solutions. In more realistic situations, such as those presented here, polynomial functions are unable to correctly capture the behavior of the Hamiltonian derivative. In fact, even a fit to a polynomial of degree 10 using the full data set shows unphysical oscillations near the integration limits (see Figure S3, Supporting Information). We have also tested some alternative functional forms (e.g., rational functions), but although reasonable, their overall performance was not as good as that of eq 18. These studies are presented in detail in section S.2 of the Supporting Information.

**4.3. Applicability Test.** Our study of different integration methods, performed above, focused on two small solutes, so as to enable simulations at a large number of intermediate values of $\lambda$. In this section, we assess whether the conclusions drawn from the analysis of the prototype systems are applicable in realistic free energy calculations involving more complex molecules. For that purpose, we attempt to compute the hydration energy of 2-methylbutanol and the hydration energy of a multifunctional compound (4-nitrophenol) using the methodology proposed above.

Previously, we have seen that the deviation in the electrostatic contribution to the free energy was very small and practically independent of the integration method down to $N = 17$ (Figure 2). The same can be said of the LJ component down to $N = 33$ (Figure 5). For that reason, we have carried out simulations for 2-methylbutanol and 4-nitrophenol using 17 points for the electrostatic component and 31 points for the LJ component, to serve as reference values. Our previous analysis showed that sufficiently precise free energies could be obtained with $N = 5$ for the electrostatic component (using the fitting function or Simpson's rule) and $N = 11$ for the LJ component (using the fitting function). Thus, we have generated reduced data sets with these values of $N$ for each respective component. The full results of the fitting procedure are given in Supporting Information, Tables S11 to S14 (including additional reduced data sets that were tested).

In Figure 6, we show the fits to the full and reduced data sets of 2-methylbutanol using eqs 7 and 18 for the electrostatic and LJ contributions, respectively. In both cases, the fits using the reduced data sets are able to provide a good description of the behavior of the Hamiltonian derivative. In Tables 2 and 3, we present the reference values for each contribution (full data set integrated using the Simpson rule) as well as the deviations from this value using the reduced sets and different integration methods. The analysis of both solutes confirms our previous conclusions based on methane and methanol—good results for the electrostatic component (error below 0.15 kJ/mol) are obtained using either the Simpson rule or the fitting function, while for the LJ component, only the fitting function is able to provide sufficiently precise free energies (error of 0.15 kJ/mol) based on the reduced data sets. The results are even more striking for 4-nitrophenol, with errors below 0.05 kJ/mol obtained using our suggested protocol, particularly considering the complexity of this multifunctional molecule. This confirms our claim that a correct choice of integration method can substantially improve the precision of solvation free energy calculations, even for complex solutes. Another way of thinking about this is to say that using our proposed integration methods one can make free energy calculations faster by a factor between 3 and 4, by reducing the necessary number of intermediate points, without a significant loss in precision.

Table 4 summarizes our results for the total hydration energy of the four solutes considered. The reference values

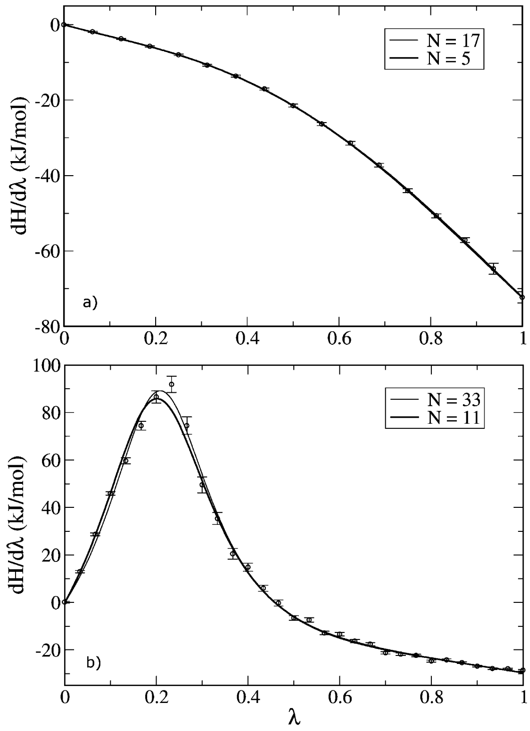


***Figure 6.*** Fits to the data for methylbutanol using the full and reduced data sets for the (a) electrostatic contribution using eq 7 and (b) Lennard-Jones contribution using eq 18.

***Table 2.*** Results (in kJ/mol) for the Two Contributions to the Hydration Energy of 2-Methylbutanol and Deviations from the Reference Value Using Different Integration Methods

| | electrostatic | Lennard-Jones |
|---|---|---|
| $\Delta G_{\mathrm{Reference}}$ | −26.86 | 9.62 |
| $\lvert\Delta G_{\mathrm{Trapezoidal}} - \Delta G_{\mathrm{Reference}}\rvert$ | 0.568 | 0.623 |
| $\lvert\Delta G_{\mathrm{Simpson}} - \Delta G_{\mathrm{Reference}}\rvert$ | 0.103 | 0.901 |
| $\lvert\Delta G_{\mathrm{Analytic}} - \Delta G_{\mathrm{Reference}}\rvert$ | 0.140 | 0.152 |

***Table 3.*** Results (in kJ/mol) for the Two Contributions to the Hydration Energy of 4-Nitrophenol and Deviations from the Reference Value Using Different Integration Methods

| | electrostatic | Lennard-Jones |
|---|---|---|
| $\Delta G_{\mathrm{Reference}}$ | −33.39 | 1.75 |
| $\lvert\Delta G_{\mathrm{Trapezoidal}} - \Delta G_{\mathrm{Reference}}\rvert$ | 0.521 | 0.365 |
| $\lvert\Delta G_{\mathrm{Simpson}} - \Delta G_{\mathrm{Reference}}\rvert$ | 0.042 | 0.509 |
| $\lvert\Delta G_{\mathrm{Analytic}} - \Delta G_{\mathrm{Reference}}\rvert$ | 0.010 | 0.044 |

***Table 4.*** Results for the Total Hydration Energy (in kJ/mol) of the Four Solutes Compared to Experimental Data[37,38]

| solute | $\Delta G_{\text{Reference}}$ | $\Delta G_{\text{Analytic}}$ | $\Delta G_{\text{Experimental}}$ |
|---|---|---|---|
| methane | 9.0 | 8.9 | 8.1 |
| methanol | −19.8 | −20.0 | −21.2 |
| 2-methylbutanol | −17.2 | −17.5 | −18.0 |
| 4-nitrophenol | −31.6 | −31.7 | −44.0 |

(from the full data sets) are compared to results obtained using reduced data sets of the recommended size (11 for LJ and 5 for electrostatic) integrated using the fitting functions. Although it is not our aim here to discuss the accuracy of the molecular model employed, it is nevertheless instructive to compare our results with experimental data. Encouragingly, our results are close to experimental values for the two simple solutes and agree very well with experimental results for 2-methylbutanol. For the case of 4-nitrophenol, the agreement is worse, which illustrates the weakness of current force-fields in predicting hydration free energies of multifunctional compounds, as discussed elsewhere.[36]

## 5. Conclusions

In this work, we have carried out a detailed analysis of the effect of the integration method on the calculation of solvation free energies using thermodynamic integration of molecular simulation data. By performing a very large number of simulations (129 for each component) at intermediate values of the coupling parameter, we have shown that the Hamiltonian derivative with respect to $\lambda$ for the electrostatic component displayed a smooth and monotonous behavior, while that for the Lennard-Jones component had a more complex shape with a prominent peak at low $\lambda$ values. For the electrostatic component, the commonly used trapezoidal rule introduces systematic errors in the free energy as the number of intermediate points decreases. However, using either Simpson's rule or a fitting polynomial of degree 4, these errors are significantly reduced, and one is able to obtain precise free energies with as few as 5 intermediate points. For the LJ component, however, both numerical integration methods show approximately similar performances, with the errors increasing substantially as the number of points decreases below about 17. We have derived a physically based fitting function that is able to provide a good description of the LJ Hamiltonian derivative throughout the entire integration interval. Analytical integration of this fitting function produces accurate free energies with as few as 11 intermediate points. It is important to notice, however, that convergence of the individual simulations is a requirement for obtaining precise free energies. Indeed, if the data set is not sufficiently converged, no integration method (including regression) will produce precise estimates. Our data were obtained using sampling times of 5 ns for each intermediate point, and convergence was checked thoroughly.

On the basis of our study of the hydration of simple solutes, we are able to recommend the following protocol for free energy calculations using thermodynamic integration: (i) for the electrostatic component, one should run simulations at 5 evenly spaced values of $\lambda$ and integrate the data using either Simpson's rule or by fitting to a quartic polynomial; (ii) for the LJ component, one should run 11 simulations at evenly spaced points, fit the data to eq 18, and calculate the free energy from the analytic integral of the fitting function, eq 20.

We have subsequently tested this protocol for more demanding cases—hydration of 2-methylbutanol and 4-nitrophenol. The results obtained confirm our previous conclusions, thus showing that the above protocol is robust and can be applied for the solvation of more complex solutes.

In summary, the use of an appropriate integration method can significantly improve the precision of free energy calculations using thermodynamic integration, for a given computational cost, or, alternatively, can make the calculations much faster for a given precision level. The integration error implicit in the TI method is commonly seen as a disadvantage of this approach relative to other methods, like thermodynamic perturbation theory. Our contribution significantly reduces this disadvantage, making TI even more competitive. We believe such improvements are required so that solvation free energy data can begin to be routinely employed in force-field parametrization and can play a more active part in drug design efforts. Although our proposed protocol and choice of fitting functions is specific to solvation free energy calculations, the principles of the method may be extended to other types of free energy calculation (e.g., potentials of mean force), with appropriate adaptations in the functional forms and in the required number of intermediate points.

**Acknowledgment.** The authors are grateful for the support provided by *Fundação para a Ciência e a Tecnologia* (FCT), Portugal, through projects FEDER/POCI/2010 and REEQ/1164/EQU/2005. N.M.G. also acknowledges his Ph.D. scholarship SFRH/BD/47822/2007 from FCT.

**Supporting Information Available:** Detailed van der Waals parameters, point charges, bond stretching, bond angle bending, and torsional force constants as well as detailed bonded and nonbonded potential parameters are provided for all compounds studied. The full data sets for the different contributions to the derivative of the Hamiltonian with respect to $\lambda$ and the full results of the analysis of the LJ and electrostatic terms are also provided. Finally, results obtained by using alternative fitting functions to the one presented in the paper for the LJ component of the free energy are illustrated. This information is available free of charge via the Internet at http://pubs.acs.org/.

## References

(1) Gilson, M.; Zhou, H. Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42.

(2) Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303* (5665), 1813–1818.

(3) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Shirts, M. R.; Dill, K. A. Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations. *J. Chem. Theory Comput.* **2009**, *5* (2), 350–358.

(4) Guthrie, J. P. A Blind Challenge for Computational Solvation Free Energies: Introduction and Overview. *J. Phys. Chem. B* **2009**, *113* (14), 4501–4507.

(5) Oostenbrink, C.; Villa, A.; Mark, A. E.; Van Gunsteren, W. F. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.* **2004**, *25* (13), 1656–1676.

(6) Kirkwood, J. G. Statistical Mechanics of Pure Fluids. *J. Chem. Phys.* **1935**, *3*, 300–313.

(7) Chipot, C.; Pohorille, A. *Free Energy Calculations - Theory and Applications in Chemistry and Biology*; Springer: Berlin, 2007.

(8) Shirts, M. R.; Pitera, J. W.; Swope, W. C.; Pande, V. S. Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins. *J. Chem. Phys.* **2003**, *119* (11), 5740–5761.

(9) Straatsma, T.; Berendsen, H. Free energy of ionic hydration: analysis of a thermodynamic integration technique to evaluate free energy differences by molecular dynamics simulations. *J. Chem. Phys.* **1988**, *89*, 5876–5886.

(10) Smith, E. J.; Bryk, T.; Haymet, A. D. J. Free energy of solvation of simple ions: molecular-dynamics study of solvation of $Cl^-$ and $Na^+$ in the ice/water interface. *J. Chem. Phys.* **2005**, *123*, 034706.

(11) Swope, W.; Andersen, H. A molecular dynamics method for calculating the solubility of gases in liquids and the hydrophobic hydration of inert-gas atoms in aqueous solution. *J. Phys. Chem.* **1984**, *88*, 6548–6556.

(12) Hummer, G.; Pratt, L. R.; Garca, A. E. Free energy of ionic hydration. *J. Phys. Chem.* **1996**, *100* (4), 1206–1215.

(13) Shyu, C.; Ytreberg, F. M. Reducing the bias and uncertainty of free energy estimates by using regression to fit thermodynamic data. *J. Comput. Chem.* **2009**, *30*, 2297–2304.

(14) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

(15) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118* (45), 11225–11236.

(16) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. The missing term in effective pair potentials. *J. Phys. Chem.* **1987**, *91* (24), 6269–6271.

(17) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, *18* (12), 1463–1472.

(18) Miyamoto, S.; Kollman, P. A. SETTLE - An Analytical Version of the SHAKE and RATTLE agorithm for Rigid Water Molecules. *J. Comput. Chem.* **1992**, *13* (8), 952–962.

(19) Garrido, N. M.; Jorge, M.; Queimada, A. J.; Economou, I. G.; Macedo, E. A. Molecular Simulation of the Hydration Gibbs Energy of Barbiturates. *Fluid Phase Equilib.* **2010**, *289*, 148−155.

(20) Lee, F. S.; Warshel, A. A local reaction field method for fast evaluation of long-range electrostatic interactions in molecular simulations. *J. Chem. Phys.* **1992**, *97* (5), 3100–3107.

(21) van Gunsteren, W.; Berendsen, H. A leap-frog algorithm for stochastic dynamics. *Mol. Simul.* **1988**, *1* (3), 173–185.

(22) Van Gunsteren, W. F.; Berendsen, H. J. C. Algorithms for Brownian Dynamics. *Mol. Phys.* **1982**, *45* (3), 637–647.

(23) Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81* (8), 3684–3690.

(24) Beuler, T. M. R.; van Schaik, R. C.; Gerber, P. R.; van Gunsteren, W. F. Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chem. Phys. Lett.* **1994**, *222*, 529–539.

(25) Pitera, J. W.; Van Gunsteren, W. F. A comparison of non-bonded scaling approaches for free energy calculations. *Mol. Simul.* **2002**, *28* (1−2), 45–65.

(26) Mobley, D. L.; Dumont, E.; Chodera, J. D.; Dill, K. A. Comparison of charge models for fixed-charge force fields: Small-molecule hydration free energies in explicit solvent. *J. Phys. Chem. B* **2007**, *111* (9), 2242–2254.

(27) Liu, D. C.; Nocedal, J. On the Limited Memory BFGS Method for Large Scale Optimization. *Math. Program.* **1989**, *45* (3), 503–528.

(28) Flyvbjerg, H.; Petersen, H. Error estimates on averages of correlated data. *J. Chem. Phys.* **1989**, *91* (1), 461–466.

(29) Press, W.; Teukolsky, S.; Vetterling, W.; Flannery, B. *Numerical recipes in C*, 2nd ed.; Cambridge University Press: Cambridge, U. K., 1992.

(30) Pierotti, R. A. A scaled particle theory of aqueous and nonaqueous solution. *Chem. Rev.* **1976**, *76* (6), 717–726.

(31) Stillinger, F. H. Structure in aqueous solutions of nonpolar solutes from the standpoint of scaled-particle theory. *J. Sol. Chem.* **1973**, *2* (2/3), 141–158.

(32) Reiss, H.; Frisch, H. L. Statistical mechanics of rigid spheres. *J. Chem. Phys.* **1959**, *31* (2), 369–380.

(33) Westergren, J.; Lindfors, L.; Hoglund, T.; Luder, K.; Nordholm, S.; Kjellander, R. In silico prediction of drug solubility: 1. Free energy of hydration. *J. Phys. Chem. B* **2007**, *111* (7), 1872–1882.

(34) Grace Software is available free of charge at http://plasma-gate.weizmann.ac.il/Grace/ (acessed October 22, 2009).

(35) Wan, S. Z.; Stote, R. H.; Karplus, M. Calculation of the aqueous solvation energy and entropy, as well as free energy, of simple polar solutes. *J. Chem. Phys.* **2004**, *121* (19), 9539–9548.

(36) Garrido, N. M.; Queimada, A. J.; Jorge, M.; Economou, I. G.; Macedo, E. A. Molecular Simulation of Absolute Hydration Gibbs Energies of Polar Compounds. Submitted for Publication, 2010.

(37) Michielan, L.; Bacilieri, M.; Kaseda, C.; Moro, S. Prediction of the Aqueous Solvation Free Energy of Organic Compounds by Using Autocorrelation of Molecular Electrostatic Potential Surface Properties Combined with Response Surface Analysis. *Bioorg. Med. Chem.* **2008**, *16* (10), 5733–5742.

(38) Cabani, S.; Gianni, P.; Mollica, V.; Lepori, L. Group Contributions to the Thermodynamic Properties of Non-Ionic Organic Solutes in Dilute Aqueous Solution. *J. Sol. Chem.* **1981**, *10* (8), 563–595.

CT900661C

JCTC Journal of Chemical Theory and Computation

# An Atomic-Orbital-Based Lagrangian Approach for Calculating Geometric Gradients of Linear Response Properties

Sonia Coriani*

*Dipartimento di Scienze Chimiche, Università degli Studi di Trieste, via L. Giorgieri 1, I-34127 Trieste, Italy and Centre for Theoretical and Computational Chemistry, University of Oslo, P.O. Box 1033, N-0315, Blindern, Norway*

Thomas Kjærgaard and Poul Jørgensen

*Lundbeck Center for Theoretical Chemistry, University of Aarhus, DK-8000 Århus C, Denmark*

Kenneth Ruud

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

Joonsuk Huh and Robert Berger

*Frankfurt Institute for Advanced Studies (FIAS), Johann Wolfgang Goethe-University, D-60438 Frankfurt am Main, Germany*

Received September 25, 2009

**Abstract:** We present a Lagrangian approach for the calculation of molecular (quadratic) response properties that can be expressed as geometric gradients of a generic linear response function, its poles, and its residues. The approach is implemented within an atomic-orbital-based formalism suitable for linear scaling at the level of self-consistent time-dependent Hartree−Fock and density functional theory. Among the properties that can be obtained using this formalism are the gradient of the frequency-dependent polarizability (e.g., Raman intensities) and that of the one-photon transition dipole moment (entering the Herzberg−Teller factors), in addition to the excited-state molecular forces required for excited-state geometry optimizations. Geometric derivatives of ground-state first-order properties (e.g., IR intensities) and excited-state first-order property expressions are also reported as byproducts of our implementation. The one-photon transition moment gradient is the first analytic implementation of the one-photon transition moment derivative at the DFT level of theory. Besides offering a simple solution to overcome phase (hence, sign) uncertainties connected to the determination of the Herzberg−Teller corrections by numerical derivatives techniques based on independent calculations, our approach also opens the possibility to determine, for example by a mixed analytic−numerical approach, the one-photon transition dipole Hessian, and thus to investigate vibronic effects beyond the linear Herzberg−Teller approximation. As an illustrative application, we report a DFT study of the vibronic fine structure of the one-photon $\tilde{X}(^1A_{1g}) - \tilde{A}(^1B_{2u})$ transition in the absorption spectrum of benzene, which is Franck−Condon-forbidden in the electric dipole approximation and hence determined by the Herzberg−Teller integrals and electronic transition dipole-moment derivatives.

## 1. Introduction

Derivatives (of electronic properties) with respect to displacements of the nuclei, for brevity denoted as geometric (property) derivatives herein, are one of the key ingredients in describing the effect of molecular vibrations on properties computed within the Born−Oppenheimer approximation, as well as selection rules for a variety of spectroscopic effects, such as Raman or infrared spectroscopy.[1−10] The geometric

* Corresponding author e-mail: coriani@units.it.

10.1021/ct900506c © 2010 American Chemical Society
Published on Web 03/02/2010

derivatives of the dynamic electric dipole polarizability determine, for instance, the intensity of the Raman spectrum[3,9] and play a fundamental role in the theoretical description of other Raman processes, like coherent anti-Stokes Raman Spectroscopy (CARS)[11] and vibrational Raman optical activity.[8,10,12] Similarly, the first geometric derivative of the transition dipole strength yields information on how the motion of the nuclei will affect the UV spectrum (or one-photon absorption, OPA) of a molecule through the so-called Herzberg−Teller (HT) contribution (vibronic coupling between different electronic states).[13−15] Finally, the first geometric derivative of the excited-state energy is the excited-state gradient, which can be used to determine and characterize the equilibrium geometry of a system in an excited electronic state.

Geometric derivatives can be obtained either numerically or analytically. Analytical approaches are more time-consuming to implement in computer codes than numerical methods, but they have the clear advantage of being numerically stable and yielding more accurate results than numerical differentiation, as well as being generally available, once implemented, for any type of system (see, e.g., the discussion in ref 16). Analytic and numerical derivative schemes can also be combined: for instance, the excited-state Hessian can be obtained from the numerical derivative of the analytic excited-state gradient, with higher numerical accuracy than would be obtained in a fully numerical second-order derivative procedure applied to the excited-state energy.

An efficient way to obtain derivatives of a property (usually the energy) is by the Lagrange multipliers technique of multivariable calculus, typically used in constrained optimization problems. The Lagrangian technique has been used in various contexts within the quantum chemistry community, and in electronic structure theory in particular.[9,17−23] A well-known example is the energy/wave-function optimization for both variational and nonvariational wave function approximations, see for instance refs 24 and 25. When implementing analytic geometry derivatives, the Lagrangian technique is used in a nontraditional way, in which the Lagrangian multipliers are treated as wave-function parameters on equal footing with the conventional wave-function parameters. The variational nature of the Lagrangian is then used to reduce the number of response equations that need to be solved. The use of the Lagrangian approach for calculating geometric derivatives was introduced by Helgaker and Jørgensen at the end of the 1980s,[16,18,20] and it has been used, for instance, to obtain the geometric derivatives of ab initio electronic energy surfaces, as well as magnetic derivatives of the energy using perturbation-dependent basis sets.[9,23,26−28] With the introduction of the quasi-energy approach to frequency-dependent response properties,[21,22] the Lagrangian method has been shown to afford efficient computational expressions for the implementation of both dynamic response functions and multiphoton transition moments at various levels of theory, in particular, the coupled-cluster[22,29,30] and, recently, time-dependent density functional[31,32] theories.

In this paper, we use a Lagrangian technique to determine the working equations for third-order molecular properties that are related to the geometric derivatives of second-order response properties—that is, frequency-dependent linear response functions, their poles, and their residues. In the specific case of the electric dipole polarizability and that of the electric-dipole transition strength, these derivatives will correspond to the electric dipole polarizability gradient, which determines the intensity of the Raman spectrum, and to the Herzberg−Teller contribution to the OPA spectrum, respectively. Moreover, as the pole of the linear response function occurs at an electronic excitation from the ground state to an excited state, its geometric first derivative automatically yields the excited-state gradient, and this in turns opens the possibility of obtaining the optimized equilibrium structure of an excited state directly from a ground-state wave function/density.

Since the starting expressions for the second-order quantities to be differentiated are taken according to the atomic orbital formulation of response theory presented in refs 33 and 34—which is based on an exponential parametrization of the atomic-orbital density matrix[25,35]—the resulting properties can be calculated at linear computational cost for sufficiently sparse matrices and can be easily parallelized, since all references to individual two-electron (derivative) integral distributions are avoided and only elementary matrix operations have to be done. The atomic orbital basis we adopt represents a convenient framework for deriving properties whose dependence on the perturbation is already contained in the orbitals. As byproducts of our derivation, we also give the expressions for both the ground- and excited-state first-order properties (e.g., the dipole moments) and the geometric gradient of the ground-state first-order properties (required, for instance, to obtain IR intensities).

Even though we here only consider geometric derivatives, the approach is quite general and has been applied, with a few modifications, to derive and implement working expressions for the magnetic derivatives of second-order properties, using London atomic orbitals to ensure gauge-origin independence.[28] Starting, for instance, from the (imaginary) electric-dipole polarizability and transition strengths, these yield the hyperpolarizability that enters the Verdet constant, and the transition strength that gives the Faraday B term of magneto-optical activity (i.e., the Faraday effect in the transparent and absorptive regions of the sample).[28,36,37] Geometric and magnetic perturbations are treated on an equal footing when using so-called perturbation-dependent basis sets since the atomic orbitals in both of these cases depend explicitly on the differentiating variable. Note, moreover, that the expressions we obtain contain, as a subset, the standard expression for the quadratic response function $\langle\langle A; B, C\rangle\rangle_{\omega,0}$ and its residues when the geometric perturbation is replaced by a generic one-electron (static) operator $C$.

The procedure we adopt is equivalent to the one used by Furche and co-workers[9,23] to obtain the excited-state gradient and vibrational Raman intensities in time-dependent density functional theory, differing only in that the derivation of Furche and co-workers is expressed in a conventional molecular-orbital basis. Thorvaldsen et al.[32] have also very recently presented a general method for the calculation of molecular properties to arbitrary order, in which the quasienergy and Lagrangian formalisms are combined to derive response functions by differentiation of the quasienergy derivative Lagrangian using the elements of the density matrix in the atomic orbital representation as variational parameters. The method has been applied to compute, at the Hartree−Fock level, the CARS spectra[11] of a series of polycyclic aromatic hydrocarbons and the vibrational (hyper)-polarizabilities (which require the geometric gradients of the electric dipole moment, electric polarizability, and electric first hyperpolarizability) of water, of three all-trans polyenes, and of three 4-dimethylaminophenylpolyene aldehydes.[38] Excited-state and (hyper)polarizability gradient implementations have thus been reported previously, while this work presents the first implementation of the analytic computation of the transition moment derivatives. We also note that the analytic evaluation of transition moment derivatives offers a simple solution to overcome phase (hence, sign) uncertainties connected to the their determination by numerical derivatives techniques based on totally independent calculations. In principle, it can also be combined with a numerical differentiation scheme, as commonly done to determine the excited-state Hessian, to yield the transition-moment second-order derivates, allowing the investigation of vibronic effects beyond the linear Herzberg−Teller approximation (note however that, in such mixed numerical−analytical schemes for transition moments, the phase uncertainties may re-emerge).

This paper is organized as follows. In the Theory section, we first define the key quantities which represent the formal background for our derivation. We then outline the Lagrange method both in general terms and in a more specific way for the properties of interest. The implemented expressions for the second-order property derivatives (third-order properties) will be given at the end of the section.

As an illustrative application, we report in the Illustrative Results section an exhaustive DFT study of the vibronic fine structure of the one-photon $\tilde{X}(^1A_{1g}) - \tilde{A}(^1B_{2u})$ transition in the absorption spectrum of benzene. This transition is Franck−Condon-forbidden in the electric dipole approximation and hence dominated by the (first-order) Herzberg−Teller integrals and electronic transition dipole-moment derivatives.[13,39]

## 2. Theory

**2.1. Ansatz: Exponential Parametrization of the Density.** We start by assuming that the wave function (or density) of the ground state is optimized for a point on the potential surface ($\mathbf{R}_0$) such that the variational condition at that point is fulfilled:[40]

$$\mathbf{E}^{[1]} = \mathbf{FDS} - \mathbf{SDF} = \mathbf{0} \tag{1}$$

where $\mathbf{E}^{[1]}$ is the matrix representation of the electronic gradient in the nonorthogonal atomic orbital (AO) basis, $\mathbf{S}$ is the AO overlap matrix, $\mathbf{D}$ is the AO density, and $\mathbf{F}$ is the Fock/Kohn−Sham matrix:

$$\mathbf{F} = \mathbf{h} + \mathbf{G}^{\mathrm{HF}}(\mathbf{D}) \tag{2}$$

In the equation above, $\mathbf{h}$ is the AO integral matrix for the one-electron part (kinetic plus nuclear attraction) of the Hamilton operator and $\mathbf{G}(\mathbf{D})$ denotes the Coulomb and exact-exchange contributions:

$$G^{\mathrm{HF}}_{\mu\nu}(\mathbf{D}) = \sum_{\rho\sigma} D_{\sigma\rho}[g_{\mu\nu\rho\sigma} - w_x g_{\mu\sigma\rho\nu}] \tag{3}$$

The scaling factor $w_x$ is equal to 1 for Hartree−Fock. In the case of Kohn−Sham theory, the scaling factor $w_x$ is zero unless a hybrid functional is used, and an additional contribution from the exchange-correlation potential must be included in the Kohn−Sham matrix,[41]

$$\mathbf{F} = \mathbf{h} + \mathbf{G}^{\mathrm{HF}}(\mathbf{D}) + \mathbf{F}^{\mathrm{xc}} \tag{4}$$

The last term in eq 4 is the derivative of the exchange-correlation functional $E_{\mathrm{xc}}[\rho]$:

$$F^{\mathrm{xc}}_{\mu\nu} = \frac{\partial E_{\mathrm{xc}}[\rho]}{\partial D_{\nu\mu}} \tag{5}$$

Expressing the density $\rho$ in the AO basis as

$$\rho(\mathbf{r}) = \sum_{\mu\nu} \chi^*_\mu(\mathbf{r})\, \chi_\nu(\mathbf{r})\, D_{\nu\mu} = \sum_{\mu\nu} \Omega_{\mu\nu}(\mathbf{r})\, D_{\nu\mu} \tag{6}$$

where $\Omega_{\mu\nu}(\mathbf{r}) = \chi^*_\mu(\mathbf{r})\, \chi_\nu(\mathbf{r})$ is the overlap distribution, and introducing the exchange-correlation potential

$$v_{\mathrm{xc}}(\mathbf{r}) = \frac{\delta E_{\mathrm{xc}}[\rho]}{\delta \rho(\mathbf{r})} \tag{7}$$

we see that

$$F^{\mathrm{xc}}_{\mu\nu} = \int \frac{\delta E_{\mathrm{xc}}[\rho]}{\delta \rho(\mathbf{r})} \frac{\partial \rho(\mathbf{r})}{\partial D_{\nu\mu}}\, \mathrm{d}\mathbf{r} = \int v_{\mathrm{xc}}(\mathbf{r})\, \Omega_{\mu\nu}(\mathbf{r})\, \mathrm{d}\mathbf{r} \tag{8}$$

The ground-state energy at $\mathbf{R}_0$ is obtained as

$$E_0 = \mathrm{Tr}\, \mathbf{hD} + \frac{1}{2}\mathrm{Tr}\, \mathbf{DG}^{\mathrm{HF}}(\mathbf{D}) + E_{\mathrm{xc}}[\rho] + h_{\mathrm{nuc}} \tag{9}$$

where $h_{\mathrm{nuc}}$ is the nuclear repulsion term.

The variational condition in the form given in eq 1—as well as the response expressions in the next sections—was derived on the basis of an exponential parametrization of the AO density matrix:[25,33,35,40]

$$\mathbf{D}(\mathbf{X}) = \exp(-\mathbf{XS})\mathbf{D}\exp(\mathbf{SX}) = \mathbf{D} + [\mathbf{D}, \mathbf{X}]_{\mathrm{S}} + \frac{1}{2}[[\mathbf{D}, \mathbf{X}]_{\mathrm{S}}, \mathbf{X}]_{\mathrm{S}} + ... \tag{10}$$

where $\mathbf{X}$ is an anti-Hermitian matrix that contains the variational parameters, with the redundant parameters projected out:

$$\mathbf{X} = \mathscr{P}(\mathbf{X}) \equiv \mathbf{P}_{\mathrm{o}}\mathbf{X}\mathbf{P}_{\mathrm{v}}^{\mathrm{T}} + \mathbf{P}_{\mathrm{v}}\mathbf{X}\mathbf{P}_{\mathrm{o}}^{\mathrm{T}} \tag{11}$$

$\mathbf{P}_{\mathrm{o}}$ and $\mathbf{P}_{\mathrm{v}}$ are projectors onto the occupied and virtual orbital spaces, respectively:

$$\mathbf{P}_{\mathrm{o}} = \mathbf{DS} \tag{12}$$

$$\mathbf{P}_{\mathrm{v}} = \mathbf{I} - \mathbf{DS} \tag{13}$$

fulfilling the idempotency ($\mathbf{P}_{\mathrm{o}}^2 = \mathbf{P}_{\mathrm{o}}$ and $\mathbf{P}_{\mathrm{v}}^2 = \mathbf{P}_{\mathrm{v}}$) and orthogonality relations ($\mathbf{P}_{\mathrm{o}}\mathbf{P}_{\mathrm{v}} = \mathbf{P}_{\mathrm{v}}\mathbf{P}_{\mathrm{o}} = \mathbf{0}$ and $\mathbf{P}_{\mathrm{o}}^{\mathrm{T}}\mathbf{S}\mathbf{P}_{\mathrm{v}} = \mathbf{P}_{\mathrm{v}}^{\mathrm{T}}\mathbf{S}\mathbf{P}_{\mathrm{o}} = \mathbf{0}$). The so-called S commutator appearing in eq 10 is defined as

$$[\mathbf{D},\mathbf{X}]_{\mathrm{S}} = \mathbf{DSX} - \mathbf{XSD} \tag{14}$$

More generally, we introduce the M commutator as

$$[\mathbf{L},\mathbf{N}]_{\mathrm{M}} = \mathbf{LMN} - \mathbf{NML} \tag{15}$$

**2.2. A Few Words on Notation.** Before proceeding, it is convenient to introduce here a more compact notation, which is repeatedly used throughout the paper. According to it, we write an element of the gradient matrix in eq 1 as[33]

$$E_m^{[1]} = \mathrm{Tr}\,\mathbf{F}[\mathbf{D},\mathbf{O}_m^{\dagger}]_{\mathrm{S}} = \mathrm{Tr}\,\mathbf{O}_m^{\dagger}(\mathbf{FDS} - \mathbf{SDF}) \equiv \mathrm{Tr}\,\mathbf{O}_m^{\dagger}\mathbf{E}^{[1]} \tag{16}$$

As a rule of thumb, we can go from the (element-wise) notation to the true matrix representation in the AO basis using

$$M_j = \mathrm{Tr}\,\mathbf{O}_j^{\dagger}\mathbf{M} \tag{17}$$

where the index $j$ indicates a $M_{\mu\nu}$ element of the matrix $\mathbf{M}$ in the AO basis. The operator $\mathbf{O}_j^{\dagger}$ and its adjoint $\mathbf{O}_j$ are defined in ref 33. We also introduce here the expansions

$$\mathbf{b}^{\omega} = \sum_m b_m^{\omega}\mathbf{O}_m; \qquad \mathbf{b}^{\omega\dagger} = \sum_m b_m^{\omega *}\mathbf{O}_m^{\dagger} \tag{18}$$

**2.3. Ansatz: AO-Based Linear Response Theory.** We consider now the expansion of the time-dependent expectation value of a time-independent operator $A$ with respect to a (periodic) perturbation $V^t = \int_{-\infty}^{+\infty} V^{\omega}\exp(-i\omega t)\,\mathrm{d}\omega$:

$$\langle A(t)\rangle = A_0 + \int \langle\langle A;V^{\omega}\rangle\rangle_{\omega}\exp(-i\omega t)\,\mathrm{d}\omega + \ldots \tag{19}$$

where $\langle\langle A;V^{\omega}\rangle\rangle_{\omega}$ is the linear response function. Assuming implicit summation over repeated indices, and introducing the symbol $B$ in place of $V^{\omega}$, the linear response function (LRF) is given by[33]

$$\langle\langle A;B\rangle\rangle_{\omega} = -A_m^{[1]}b_m^{\omega} = -\boldsymbol{A}^{[1]\dagger}\boldsymbol{b}^{\omega} \equiv -\mathrm{Tr}(\mathbf{A}^{[1]\dagger}\mathbf{b}^{\omega}) = + \mathrm{Tr}\mathbf{A}[\mathbf{b}^{\omega},\mathbf{D}]_{\mathrm{S}} \tag{20}$$

where the elements $b_m^{\omega}$ of the response "vector" $\boldsymbol{b}^{\omega}$ (matrix $\mathbf{b}^{\omega}$) are obtained from the solution of the linear response equation

$$(E_{mn}^{[2]} - \omega S_{mn}^{[2]})b_n^{\omega} = B_m^{[1]} \tag{21}$$

or, in supermatrix notation,[28]

$$(\boldsymbol{E}^{[2]} - \omega\boldsymbol{S}^{[2]})\boldsymbol{b}^{\omega} = \boldsymbol{B}^{[1]} \tag{22}$$

and where

$$A_m^{[1]} = -\mathrm{Tr}\,\mathbf{A}[\mathbf{O}_m,\mathbf{D}]_{\mathrm{S}} \tag{23}$$

$$\mathbf{A}^{[1]} = \mathbf{SDA}^{\dagger} - \mathbf{A}^{\dagger}\mathbf{DS} \tag{24}$$

is the *property gradient* relative to the $A$ operator (whose expectation value is perturbed). The right-hand side of the linear response equation is the *property gradient* relative to the external perturbation described by the $V^{\omega}$ ($\equiv B$) operator,

$$B_m^{[1]} = \mathrm{Tr}\,\mathbf{B}[\mathbf{D},\mathbf{O}_m^{\dagger}]_{\mathrm{S}} = \mathrm{Tr}\,\mathbf{O}_m^{\dagger}(\mathbf{BDS} - \mathbf{SDB}) \tag{25}$$

$$\mathbf{B}^{[1]} = \mathbf{BDS} - \mathbf{SDB} \tag{26}$$

$\boldsymbol{E}^{[2]}$ and $\boldsymbol{S}^{[2]}$ are the generalized electronic Hessian and metric matrices in the AO basis:[33]

$$E_{mn}^{[2]} = \mathrm{Tr}\,\mathbf{F}[[\mathbf{O}_n,\mathbf{D}]_{\mathrm{S}},\mathbf{O}_m^{\dagger}]_{\mathrm{S}} + \mathrm{Tr}\,\mathbf{G}([\mathbf{O}_n,\mathbf{D}]_{\mathrm{S}})[\mathbf{D},\mathbf{O}_m^{\dagger}]_{\mathrm{S}} \tag{27}$$

$$S_{mn}^{[2]} = \mathrm{Tr}\,\mathbf{O}_m^{\dagger}\mathbf{S}[\mathbf{D},\mathbf{O}_n]_{\mathrm{S}}\mathbf{S} \equiv -\mathrm{Tr}\,\mathbf{O}_m^{\dagger}\mathbf{S}[\mathbf{O}_n,\mathbf{D}]_{\mathrm{S}}\mathbf{S} \tag{28}$$

where we define

$$\mathbf{G}(\mathbf{M}) = \mathbf{G}^{\mathrm{HF}}(\mathbf{M}) + \mathbf{G}^{\mathrm{xc}}(\mathbf{M}) \tag{29}$$

and introduce the matrix[41]

$$G_{\mu\nu}^{\mathrm{xc}}(\mathbf{M}) = \sum_{\rho\sigma} M_{\sigma\rho}\int \frac{\delta^2 E^{\mathrm{xc}}}{\delta\rho(\mathbf{s})\,\delta\rho(\mathbf{r})}\Omega_{\mu\nu}(\mathbf{r})\,\Omega_{\rho\sigma}(\mathbf{s})\,\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{s} \tag{30}$$

Note that the elements of the (super)matrix $\boldsymbol{E}^{[2]}$ are here defined so that $\boldsymbol{E}^{[2]}$ is positive definite.

The linear response function has poles whenever the frequency $\omega$ is equal to an excitation energy $\omega_f$. The excitation energies $\omega_f$ and excitation vectors $\boldsymbol{b}^f$ (matrices $\mathbf{b}^f$) are obtained from the solution of the generalized eigenvalue equation

$$(\boldsymbol{E}^{[2]} - \omega_f\boldsymbol{S}^{[2]})\boldsymbol{b}^f = \mathbf{0} \tag{31}$$

The corresponding residue of the linear response function can be shown to be[34]

$$\lim_{\omega\to\omega_f}(\omega - \omega_f)\langle\langle A;B\rangle\rangle_{\omega} = (\boldsymbol{A}^{[1]\dagger}\boldsymbol{b}^f)(\boldsymbol{b}^{f\dagger}\boldsymbol{B}^{[1]}) = \mathrm{Tr}(\mathbf{A}^{[1]\dagger}\mathbf{b}^f)\,\mathrm{Tr}(\mathbf{b}^{f\dagger}\mathbf{B}^{[1]}) \tag{32}$$

Note that the excited-state vector is normalized on the generalized metric matrix $\boldsymbol{S}^{[2]}$, that is

$$\boldsymbol{b}^{f\dagger}\boldsymbol{S}^{[2]}\boldsymbol{b}^f = 1 \tag{33}$$

and that $\boldsymbol{S}^{[2]}$ is *not* positive definite.

**2.4. Construction of the Lagrangian and Variational Condition.** One obvious way to obtain our third-order properties would be a straightforward differentiation of the second-order property expressions previously given. However, such an approach would automatically imply that derivatives of (either or both) the linear response vectors ($\boldsymbol{b}^{\omega}$) and the eigenvectors ($\boldsymbol{b}^f$) would be required. Hence, additional equations depending (often in a rather complicated fashion) on the number of external perturbations (up to $3N$ for each component) should be solved. Such an approach has an evident drawback when dealing with properties of large systems, as the number of equations to be solved would quickly become too large to be handled.

Alternatively, computationally efficient expressions for the third-order molecular properties can be obtained using a Lagrangian technique.[16,17,19] For each second-order property (component) $\mathcal{O}$ we want to differentiate—either a linear response function, a transition moment, or an excitation energy—we construct a Lagrangian function (component) $\mathscr{L}$, adding to the second-order property in question the appropriate constraint equations that must be satisfied, each multiplied by a set of Lagrange multipliers:

$$\mathscr{L}(\mathbf{R},\boldsymbol{\lambda},\bar{\boldsymbol{\lambda}},\mathbf{X},\bar{\mathbf{X}}) = \mathscr{C}(\mathbf{R},\boldsymbol{\lambda},\mathbf{X}) + \sum_m \bar{\lambda}_m \mathscr{P}_m(\mathbf{R},\boldsymbol{\lambda},\mathbf{X}) + \sum_n \bar{X}_m \mathscr{O}_n(\mathbf{R},\mathbf{X}) \quad (34)$$

where $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\mathbf{R})$ collectively represent the "property parameters" (for instance, the response and excitation vectors), $\mathbf{X} = \mathbf{X}(\mathbf{R})$ are the "orbital parameters" [see eq 11], $\bar{\boldsymbol{\lambda}}(\mathbf{R})$ are the Lagrange multipliers connected to the property parameters (or property constraints $\mathscr{P} = 0$), and $\bar{\mathbf{X}}(\mathbf{R})$ are the Lagrange multipliers related to the orbital parameters (or orbital constraints $\mathscr{O} = 0$). The variable $\mathbf{R}$ collectively indicates the spatial coordinates of the nuclei. Note the dependence of the property constraint equations on both the spatial coordinates of the nuclei $\mathbf{R}$, the property parameters $\boldsymbol{\lambda}$, *and* the orbital parameters $\mathbf{X}$.

Next, we make the Lagrangian fully variational, imposing its stationarity with respect to any variation in both property/orbital parameters and Lagrange multipliers ($\forall i$, $\mathbf{R}$)

$$\frac{\partial \mathscr{L}(\mathbf{R},\boldsymbol{\lambda},\bar{\boldsymbol{\lambda}},\mathbf{X},\bar{\mathbf{X}})}{\partial \bar{\lambda}_i} = 0 \Leftrightarrow \mathscr{P}_i(\mathbf{R},\boldsymbol{\lambda},\mathbf{X}) = 0 \quad (35a)$$

$$\frac{\partial \mathscr{L}(\mathbf{R},\boldsymbol{\lambda},\bar{\boldsymbol{\lambda}},\mathbf{X},\bar{\mathbf{X}})}{\partial \lambda_i} = 0 \Leftrightarrow \frac{\partial \mathscr{C}(\mathbf{R},\boldsymbol{\lambda},\mathbf{X})}{\partial \lambda_i} + \sum_j \bar{\lambda}_j \frac{\partial \mathscr{P}_j(\mathbf{R},\boldsymbol{\lambda},\mathbf{X})}{\partial \lambda_i} = 0 \quad (35b)$$

$$\frac{\partial \mathscr{L}(\mathbf{R},\boldsymbol{\lambda},\bar{\boldsymbol{\lambda}},\mathbf{X},\bar{\mathbf{X}})}{\partial \bar{X}_i} = 0 \Leftrightarrow \mathscr{O}_i(\mathbf{R},\mathbf{X}) = 0 \quad (35c)$$

$$\frac{\partial \mathscr{L}(\mathbf{R},\boldsymbol{\lambda},\bar{\boldsymbol{\lambda}},\mathbf{X},\bar{\mathbf{X}})}{\partial X_i} = 0 \Leftrightarrow \frac{\partial \mathscr{C}(\mathbf{R},\boldsymbol{\lambda},\mathbf{X})}{\partial X_i} + \sum_j \bar{\lambda}_j \frac{\partial \mathscr{P}_j(\mathbf{R},\boldsymbol{\lambda},\mathbf{X})}{\partial X_i} + \sum_j \bar{X}_j \frac{\partial \mathscr{O}_j(\mathbf{R},\mathbf{X})}{\partial X_i} = 0 \quad (35d)$$

Equations 35a and 35c simply correspond to the constraint equations that determine the property parameters ($\boldsymbol{\lambda}$) and orbital parameters ($\mathbf{X}$), respectively. Equation 35b can be used to determine the property Lagrange multipliers ($\bar{\boldsymbol{\lambda}}$), whereas eq 35d affords the orbital Lagrange multipliers ($\bar{\mathbf{X}}$).

Finally, we obtain the properties of interest from the derivative of the Lagrangian with respect to the displacements of the nuclei. For the values of the parameters that satisfy eqs 35a to 35d, it yields

$$\left.\frac{\mathrm{d}\mathscr{L}(\mathbf{R},\boldsymbol{\lambda},\bar{\boldsymbol{\lambda}},\mathbf{X},\bar{\mathbf{X}})}{\mathrm{d}R_\beta}\right|_{\mathbf{R}_0} = \left.\frac{\partial\mathscr{L}(\mathbf{R},\boldsymbol{\lambda},\bar{\boldsymbol{\lambda}},\mathbf{X},\bar{\mathbf{X}})}{\partial R_\beta}\right|_{\mathbf{R}_0} = \left.\frac{\partial \mathscr{C}(\mathbf{R},\boldsymbol{\lambda},\mathbf{X})}{\partial R_\beta}\right|_{\mathbf{R}_0} + \sum_i \bar{\lambda}_i \left.\frac{\partial \mathscr{P}_i(\mathbf{R},\boldsymbol{\lambda},\mathbf{X})}{\partial R_\beta}\right|_{\mathbf{R}_0} + \sum_i \bar{X}_i \left.\frac{\partial \mathscr{O}_i(\mathbf{R},\mathbf{X})}{\partial R_\beta}\right|_{\mathbf{R}_0} \equiv \left.\frac{\mathrm{d}\mathscr{C}(\mathbf{R},\boldsymbol{\lambda},\mathbf{X})}{\mathrm{d}R_\beta}\right|_{\mathbf{R}_0} \quad (36)$$

(with $R_\beta \in \mathbf{R}$ indicating one specific spatial coordinate of the nuclei). Third-order molecular properties are thus conveniently formulated as derivatives of the variational linear-response property Lagrangian. Using the Lagrangian technique, we thus replace the calculation of the parameters' response with respect to each perturbation $R_\beta$ with the calculation of the Lagrangian multipliers, that is, one additional set of equations (eq 35b) for each $\lambda_i$ independent of the number of perturbations (and similarly for the orbital parameters $X_i$).

In practice, the solution of eqs 35a−35d, and consequent calculation of the properties according to eq 36 and its higher-order analogs, is carried out within a variational perturbation theory approach[16] by expanding the parameters in order of the (external) perturbation. When the expansions are inserted in the expressions for the property or property Lagrangian, variational conditions for each order of the perturbation are obtained in place of one condition for each value of the field.[16] In this way, molecular properties are obtained in accordance with Wigner's $2n + 1$ rule for the parameters (the parameter response to order $n$ determines the property derivatives to order $2n + 1$), as well as the stronger $2n + 2$ rule for the multipliers.[16] Due to the $(2n + 1)$ rule, and since the properties considered here are first-order properties with respect to the displacement of the nuclei, we only need to solve the above equations through zeroth order in the displacements.

We now go into detail and report explicit expressions for the various quantities entering eq 36 for the properties of interest to us.

*2.4.1. The Linear Response Function Lagrangian.* For the linear response function, the property Lagrangian is

$$\begin{aligned}\mathscr{L}^{\alpha} &= -A^{[1]}_m b^{\omega}_m + \bar{\lambda}^*_m (E^{[2]}_{mj} b^{\omega}_j - \omega S^{[2]}_{mj} b^{\omega}_j - B^{[1]}_m) - \bar{X}^*_m E^{[1]}_m \\ &\equiv -\boldsymbol{A}^{[1]\dagger}\boldsymbol{b}^{\omega} + \bar{\boldsymbol{\lambda}}^{\dagger}(\boldsymbol{E}^{[2]}\boldsymbol{b}^{\omega} - \omega \boldsymbol{S}^{[2]}\boldsymbol{b}^{\omega} - \boldsymbol{B}^{[1]}) - \bar{\boldsymbol{X}}^{\dagger}\boldsymbol{E}^{[1]}\end{aligned} \quad (37)$$

where the second equality is again given in a supermatrix notation. It is apparent that the parameter constraint equation corresponds to the linear response equation determining $\boldsymbol{b}^{\omega}$. The orbital constraint equation in eq 34 corresponds to the optimization condition that determines the orbital parameters $\mathbf{X}$, already given in eqs 16 and 1.

*2.4.2. The Residue Lagrangian.* For the residue (i.e., the one-photon transition moment),

$$\mathscr{L}^{S} = A^{[1]}_m b^f_m - \bar{\lambda}^*_m (E^{[2]}_{mj} b^f_j - \omega_f S^{[2]}_{mj} b^f_j) - \bar{\omega}(b^{f*}_m S^{[2]}_{mj} b^f_j - 1) - \bar{X}^*_m E^{[1]}_m \equiv \boldsymbol{A}^{[1]\dagger}\boldsymbol{b}^f - \bar{\boldsymbol{\lambda}}^{\dagger}(\boldsymbol{E}^{[2]}\boldsymbol{b}^f - \omega_f \boldsymbol{S}^{[2]}\boldsymbol{b}^f) - \bar{\omega}(\boldsymbol{b}^{f\dagger}\boldsymbol{S}^{[2]}\boldsymbol{b}^f - 1) - \bar{\boldsymbol{X}}^{\dagger}\boldsymbol{E}^{[1]} \quad (38)$$

Note that in this case we have two "parameter" constraint equations, namely, the generalized eigenvalue equation for the excited-state vector $\boldsymbol{b}^f$, eq 31, and the orthonormality condition on the same excited-state vector, eq 33. The orbital constraint equation is obviously the same as for the linear response function Lagrangian.

*2.4.3. The Excited State Energy Lagrangian.* Last, for the excited-state energy, $E_f = E_0 + \omega_f$, we have

$$\begin{aligned}\mathscr{L}^{E_f} &= E_0 + b_m^{f*}E_{mj}^{[2]}b_j^f - \bar{\omega}(b_m^{f*}S_{mj}^{[2]}b_j^f - 1) - \bar{X}_m E_m^{[1]}\\ &\equiv E_0 + \boldsymbol{b}^{f\dagger}\boldsymbol{E}^{[2]}\boldsymbol{b}^f - \bar{\omega}(\boldsymbol{b}^{f\dagger}\boldsymbol{S}^{[2]}\boldsymbol{b}^f - 1) - \bar{\boldsymbol{X}}^{\dagger}\boldsymbol{E}^{[1]}\end{aligned} \tag{39}$$

where the excitation frequency $\omega_f$ was rewritten as

$$\omega_f = b_m^{f*}E_{mj}^{[2]}b_j^f \equiv \boldsymbol{b}^{f\dagger}\boldsymbol{E}^{[2]}\boldsymbol{b}^f \tag{40}$$

and the orthonormalization condition on the excited-state vectors in eq 33 was used as a unique constraint equation.

*2.4.4. The Variational Conditions for the LRF Lagrangian.* It is instructive to analyze the specific outcome of the application of the variational conditions with respect to the parameters and multipliers on the three Lagrangian functions above. Thus, for the linear response function

$$\frac{\partial\mathscr{L}}{\partial\bar{X}_j^*} = 0 \Leftrightarrow E_j^{[1]} = 0 \tag{41a}$$

$$\frac{\partial\mathscr{L}}{\partial\bar{\lambda}_j^*} = 0 \Leftrightarrow E_{jl}^{[2]}b_l^{\omega} - \omega S_{jl}^{[2]}b_l^{\omega} = B_j^{[1]} \tag{41b}$$

$$\frac{\partial\mathscr{L}}{\partial b_j^{\omega}} = 0 \Leftrightarrow A_j^{[1]} = \bar{\lambda}_m^*(E_{mj}^{[2]} - \omega S_{mj}^{[2]}) \tag{41c}$$

$$\begin{aligned}\frac{\partial\mathscr{L}}{\partial X_j} = 0 \Leftrightarrow -\frac{\partial A_m^{[1]}}{\partial X_j}b_m^{\omega} + \bar{\lambda}_m^*\frac{\partial E_{ml}^{[2]}}{\partial X_j}b_l^{\omega} - \omega\bar{\lambda}_m^*\frac{\partial S_{ml}^{[2]}}{\partial X_j}b_l^{\omega} - \\ \bar{\lambda}_m^*\frac{\partial B_m^{[1]}}{\partial X_j} = \bar{X}_m^*\frac{\partial E_m^{[1]}}{\partial X_j}\end{aligned} \tag{41d}$$

It can be easily recognized that eq 41c corresponds to a *transposed* linear response equation, as used for instance to determine the $\boldsymbol{a}^{\omega}$ (or $N^A(\omega)$) vector of the quadratic response function,[41-43] that is, $\bar{\boldsymbol{\lambda}}^{\dagger} = \boldsymbol{a}^{\omega\dagger}$. Since our response solver[34] performs the linear transformation from the right, the $\boldsymbol{\lambda}^{\dagger}$ multipliers are instead obtained by solving

$$(\boldsymbol{E}^{[2]} - \omega\boldsymbol{S}^{[2]})\bar{\boldsymbol{\lambda}} = \boldsymbol{A}^{[1]} \tag{42}$$

and then *taking the adjoint* of the resulting vector (matrix) $\bar{\boldsymbol{\lambda}}$.

The equation that determines the orbital Lagrangian multipliers, eq 41d, will be discussed in section 2.5 together with the corresponding ones from the transition-moment and excitation-energy Lagrangians.

*2.4.5. The Variational Conditions for the Residue Lagrangian.* For the transition-moment Lagrangian, we have

$$\frac{\partial\mathscr{L}}{\partial\bar{X}_j^*} = 0 \Leftrightarrow E_j^{[1]} = 0 \tag{43a}$$

$$\frac{\partial\mathscr{L}}{\partial\bar{\omega}} = 0 \Leftrightarrow b_m^{f*}S_{mj}^{[2]}b_j^f = 1 \tag{43b}$$

$$\frac{\partial\mathscr{L}}{\partial\bar{\lambda}_j^*} = 0 \Leftrightarrow E_{ji}^{[2]}b_i^f - \omega_f S_{ji}^{[2]}b_i^f = 0 \tag{43c}$$

$$\frac{\partial\mathscr{L}}{\partial b_j^f} = 0 \Leftrightarrow A_j^{[1]} = \bar{\lambda}_m^*(E_{mj}^{[2]} - \omega_f S_{mj}^{[2]}) + 2\bar{\omega}b_m^{f*}S_{mj}^{[2]} \tag{43d}$$

$$\begin{aligned}\frac{\partial\mathscr{L}}{\partial X_j} = 0 \Leftrightarrow \frac{\partial A_m^{[1]}}{\partial X_j}b_m^f - \bar{\lambda}_m^*\frac{\partial E_{ml}^{[2]}}{\partial X_j}b_l^f + \omega_f\bar{\lambda}_m^*\frac{\partial S_{ml}^{[2]}}{\partial X_j}b_l^f - \\ \bar{\omega}b_m^{f*}\frac{\partial S_{ml}^{[2]}}{\partial X_j}b_l^f = \bar{X}_m^*\frac{\partial E_m^{[1]}}{\partial X_j}\end{aligned} \tag{43e}$$

Equations 43c and 43e are clearly the eigenvalue equation and orbital parameter equation, respectively.

Equation 43d requires special attention. Similar to what was done for the linear response function, we solve the adjoint equation:

$$(\boldsymbol{E}^{[2]} - \omega_f\boldsymbol{S}^{[2]})\bar{\boldsymbol{\lambda}} + 2\bar{\omega}\boldsymbol{S}^{[2]}\boldsymbol{b}^f = \boldsymbol{A}^{[1]} \tag{44}$$

This is a linear-response equation, though for a frequency $\omega$ equal to an excitation frequency $\omega_f$. Since the excitation energy is a pole of the resolvent matrix ($\boldsymbol{E}^{[2]} - \omega_f\boldsymbol{S}^{[2]}$), eq 43d is divergent with the solution vector $\bar{\boldsymbol{\lambda}}$ having an undefined component along the excitation vector $\boldsymbol{b}^f$. By multiplying from the left with $\boldsymbol{b}^{f\dagger}$, we decouple $\bar{\omega}$ from $\bar{\lambda}$ and determine the value of the multiplier $\bar{\omega}$

$$\bar{\omega} = \frac{1}{2}\boldsymbol{b}^{f\dagger}\boldsymbol{A}^{[1]} \tag{45}$$

since $\boldsymbol{b}^{f\dagger}$ ($\boldsymbol{E}^{[2]} - \omega_f\boldsymbol{S}^{[2]}$) $= \boldsymbol{0}$. Multiplying with the orthogonal complement of $\boldsymbol{b}^{f\dagger}$, which is represented by the (projection) matrix

$$\boldsymbol{P}_f^{\dagger} = \boldsymbol{I} - \boldsymbol{S}^{[2]}\boldsymbol{b}^f\boldsymbol{b}^{f\dagger} \tag{46}$$

gives

$$\boldsymbol{P}_f^{\dagger}[(\boldsymbol{E}^{[2]} - \omega_f\boldsymbol{S}^{[2]})\bar{\boldsymbol{\lambda}}] = \boldsymbol{P}_f^{\dagger}\boldsymbol{A}^{[1]} \tag{47}$$

If we now partition the solution vector $\bar{\boldsymbol{\lambda}}$ as

$$\bar{\boldsymbol{\lambda}} = \boldsymbol{P}_f\bar{\boldsymbol{\lambda}} + \gamma\boldsymbol{b}^f \tag{48}$$

with

$$\boldsymbol{P}_f = \boldsymbol{I} - \boldsymbol{b}^f\boldsymbol{b}^{f\dagger}\boldsymbol{S}^{[2]} \tag{49}$$

and introduce it in the response equation eq 44, we are left with the well-defined, nondivergent response equation

$$\boldsymbol{P}_f^{\dagger}\{(\boldsymbol{E}^{[2]} - \omega_f\boldsymbol{S}^{[2]})\boldsymbol{P}_f(\bar{\boldsymbol{\lambda}})\} = \boldsymbol{P}_f^{\dagger}(\boldsymbol{A}^{[1]}) \tag{50}$$

since the contribution along $\boldsymbol{b}^f$ automatically vanishes because of eq 31.

In practice, eq 50 is solved by means of an iterative procedure based on trial vectors, and we need to ensure that the solution vector is kept orthogonal to the excitation vector at each step in the iterative procedure.

Since our solver[34] exploits a paired structure, where the trial vectors are normalized and orthogonalized against each other in a standard Euclidian way, the basis of trial vectors is chosen as a $2 + 2n$ basis, where the first two vectors are always chosen as the excitation vector $\boldsymbol{b}^f$, and its paired counterpart $\boldsymbol{b}^{-f}$. The remaining $2n$ vectors are generated as in the standard procedure,[34] but with the additional requirement that they are always kept orthogonal, in terms of an $\boldsymbol{S}^{[2]}$ norm, to both $\boldsymbol{b}^f$ and $\boldsymbol{b}^{-f}$

$$\boldsymbol{b}_i = (\boldsymbol{I} - \boldsymbol{b}^f\boldsymbol{b}^{f\dagger}\boldsymbol{S}^{[2]} - \boldsymbol{b}^{-f}\boldsymbol{b}^{-f\dagger}\boldsymbol{S}^{[2]})\boldsymbol{b}_i;\ \forall i = 3, ..., 2n + 2 \tag{51}$$

We refer to ref 28 for a detailed discussion of the algorithm.

*2.4.6. The Variational Conditions for the Excited-State Energy Lagrangian.* Finally, for the excited-state energy Lagrangian,

$$\frac{\partial \mathscr{L}}{\partial \bar{X}_j^*} = 0 \Leftrightarrow E_j^{[1]} = 0 \tag{52a}$$

$$\frac{\partial \mathscr{L}}{\partial \bar{\omega}} = 0 \Leftrightarrow b_j^{f*}S_{ji}^{[2]}b_i^f - 1 = 0 \tag{52b}$$

$$\frac{\partial \mathscr{L}}{\partial b_i^f} = 0 \Leftrightarrow 2(E_{ji}^{[2]}b_i^f - \bar{\omega}S_{ji}^{[2]}b_i^f) = 0 \tag{52c}$$

$$\frac{\partial \mathscr{L}}{\partial X_j} = 0 \Leftrightarrow \frac{\partial E_0}{\partial X_j} + b_m^f\frac{\partial E_{ml}^{[2]}}{\partial X_j}b_l^f - \bar{\omega}b_m^{f*}\frac{\partial S_{ml}^{[2]}}{\partial X_j}b_l^f = \bar{X}_m^*\frac{\partial E_m^{[1]}}{\partial X_j} \tag{52d}$$

which allows us to identify $\bar{\omega} = \omega_f$.

**2.5. The Response Equations for "Orbital" Lagrange Multipliers $\bar{\mathbf{X}}$.** For the determination of the orbital Lagrange multipliers in the three cases, we need to solve eqs 41d, 43e, and 52d, respectively. They involve the derivatives with respect to each orbital parameter $X_i$ of the elements of the generalized Hessian ($\boldsymbol{E}^{[2]}$) and metric ($\boldsymbol{S}^{[2]}$) matrices, as well as of the property gradients ($\boldsymbol{A}^{[1]\dagger}$ and/or $\boldsymbol{B}^{[1]}$) and of the electronic gradient ($\boldsymbol{E}^{[1]}$). These can be shown to correspond to

$$\frac{\partial E_m^{[1]}}{\partial X_j} \Rightarrow E_{mj}^{[2]} \tag{53}$$

$$\frac{\partial B_m^{[1]}}{\partial X_j} \Rightarrow B_{mj}^{[2]} = -\mathrm{Tr}\,\mathbf{B}[[\mathbf{O}_j, \mathbf{D}]_\mathrm{S}, \mathbf{O}_m^\dagger]_\mathrm{S} \tag{54}$$

$$\frac{\partial A_m^{[1]}}{\partial X_j} \Rightarrow A_{mj}^{[2]} = \mathrm{Tr}\,\mathbf{A}[\mathbf{O}_m, [\mathbf{O}_j, \mathbf{D}]_\mathrm{S}]_\mathrm{S} \tag{55}$$

$$\begin{aligned}\frac{\partial E_{mn}^{[2]}}{\partial X_j} \Rightarrow E_{mnj}^{[3]} = {}& -\mathrm{Tr}\,\mathbf{F}[[\mathbf{O}_n, [\mathbf{O}_j, \mathbf{D}]_\mathrm{S}], \mathbf{O}_m^\dagger]_\mathrm{S} - \\ & \mathrm{Tr}\,\mathbf{G}([\mathbf{O}_j, \mathbf{D}]_\mathrm{S})[[\mathbf{O}_n, \mathbf{D}]_\mathrm{S}, \mathbf{O}_m^\dagger]_\mathrm{S} - \\ & \mathrm{Tr}\,\mathbf{G}([\mathbf{O}_n, \mathbf{D}]_\mathrm{S})[[\mathbf{O}_j, \mathbf{D}]_\mathrm{S}, \mathbf{O}_m^\dagger]_\mathrm{S} - \\ & \mathrm{Tr}\,\mathbf{G}([\mathbf{O}_n, [\mathbf{O}_j, \mathbf{D}]_\mathrm{S}]_\mathrm{S})[\mathbf{D}, \mathbf{O}_m^\dagger]_\mathrm{S} - \\ & \quad \mathrm{Tr}\,\mathbf{T}^\mathrm{xc}([\mathbf{O}_n, \mathbf{D}]_\mathrm{S},\ [\mathbf{O}_j, \mathbf{D}]_\mathrm{S})[\mathbf{D}, \mathbf{O}_m^\dagger]_\mathrm{S}\end{aligned} \tag{56}$$

$$\frac{\partial S_{mn}^{[2]}}{\partial X_j} \Rightarrow S_{mnj}^{[3]} = -\mathrm{Tr}\,\mathbf{O}_m^\dagger\mathbf{S}[\mathbf{O}_n, [\mathbf{D}, \mathbf{O}_j]_\mathrm{S}]_\mathrm{S}\mathbf{S} \tag{57}$$

where we have taken advantage of the rule

$$\frac{\partial \mathbf{D}}{\partial X_j} = -[\mathbf{O}_j, \mathbf{D}]_\mathrm{S} \tag{58}$$

which stems from the exponential parametrization of the density, and introduced the matrix $\mathbf{T}^{\mathrm{xc}}$[41]

$$T_{\mu\nu}^\mathrm{xc}(\mathbf{N}, \mathbf{M}) = \sum_{\rho\sigma\eta\varepsilon} M_{\sigma\rho}N_{\varepsilon\eta}\int \Omega_{\eta\varepsilon}(\mathbf{t})\ \Omega_{\rho\sigma}(\mathbf{s})\ \Omega_{\mu\nu}(\mathbf{r})\frac{\delta^2 v_\mathrm{xc}(\mathbf{r})}{\delta\rho(\mathbf{s})\ \delta\rho(\mathbf{t})}\mathrm{d}\mathbf{r}\ \mathrm{d}\mathbf{s}\ \mathrm{d}\mathbf{t} \tag{59}$$

Using the differentiated matrices introduced above allows us to rewrite eqs 41d, 43e, and 52d, in the form

$$\bar{X}_m^*E_{mj}^{[2]} = \eta_j^*; \qquad \bar{\boldsymbol{X}}^\dagger\boldsymbol{E}^{[2]} = \boldsymbol{\eta}^\dagger \tag{60}$$

where the explicit values of the elements of the right-hand-side vector (matrix) $\boldsymbol{\eta}^\dagger$ vary according to the property we are differentiating. For the linear-response-function derivative, we get

$$\eta_j^* = -b_m^\omega A_{mj}^{[2]} + a_m^{\omega*}E_{mnj}^{[3]}b_n^\omega - \omega a_m^{\omega*}S_{mnj}^{[3]}b_n^\omega - a_m^{\omega*}B_{mj}^{[2]} \tag{61}$$

whereas for the residue (transition moment) derivative

$$\eta_j^* = b_m^fA_{mj}^{[2]} - a_m^{\omega_f*}E_{mnj}^{[3]}b_n^f + \omega_f(a_m^{\omega_f*}S_{mnj}^{[3]}b_n^{f*}) - \bar{\omega}(b_m^{f*}S_{mnj}^{[3]}b_n^f) \tag{62}$$

with $\bar{\omega}$ as in eq 45. Finally, for the excitation energy derivative

$$\eta_j^* = b_m^{f*}E_{mnj}^{[3]}b_n^f - \omega_f(b_m^{f*}S_{mnj}^{[3]}b_n^f) \tag{63}$$

where $\partial E_0/\partial X_j = 0$ because it corresponds to the optimization condition, eq 16. As previously done for the $\bar{\boldsymbol{\lambda}}$, we recast the equation in the adjoint form

$$\boldsymbol{E}^{[2]}\bar{\boldsymbol{X}} = \boldsymbol{\eta} \tag{64}$$

and solve it by reusing the iterative, linear-scaling, response solver (with transformation on the right index) of ref 34 with a modified right-hand-side matrix.

**2.6. The Generalized Property Gradient with Respect to Displacements of the Nuclei.** Once all of the parameters and Lagrange multipliers have been determined, we can turn our attention to the third-order properties, which are obtained as derivatives of the Lagrangians with respect to displacements of the nuclei, as indicated in eq 36. The final geometric derivative of the linear response function reads

$$\frac{\mathrm{d}\mathcal{C}}{\mathrm{d}R_\beta} = -\frac{\partial A_m^{[1]}}{\partial R_\beta} b_m^\omega + a_m^{\omega *}\frac{\partial E_{mn}^{[2]}}{\partial R_\beta} b_n^\omega - \omega a_m^{\omega *}\frac{\partial S_{mn}^{[2]}}{\partial R_\beta} b_n^\omega - a_m^{\omega *}\frac{\partial B_m^{[1]}}{\partial R_\beta} - \bar{X}_m^* \frac{\partial E_m^{[1]}}{\partial R_\beta} \qquad (65)$$

The transition moment gradient is

$$\frac{\mathrm{d}\mathcal{C}}{\mathrm{d}R_\beta} = \frac{\partial A_m^{[1]}}{\partial R_\beta} b_m^f - \frac{1}{2}(A_l^{[1]} b_l^f) b_m^{f*}\frac{\partial S_{mn}^{[2]}}{\partial R_\beta} b_n^f - a_m^{\omega_f *}\left(\frac{\partial E_{mn}^{[2]}}{\partial R_\beta} - \omega_f \frac{\partial S_{mn}^{[2]}}{\partial R_\beta} - \frac{\partial \omega_f}{\partial R_\beta} S_{mn}^{[2]}\right) b_n^f - \bar{X}_m^* \frac{\partial E_m^{[1]}}{\partial R_\beta} \qquad (66)$$

where

$$\frac{\partial \omega_f}{\partial R_\beta} = b_m^{f*}\frac{\partial E_{mn}^{[2]}}{\partial R_\beta} b_n^f - \omega_f b_m^{f*}\frac{\partial S_{mn}^{[2]}}{\partial R_\beta} b_n^f \qquad (67)$$

Note, however, the term including $\partial\omega_f/\partial R_\beta$ in the residue gradient vanishes since $a_m^{\omega_f}{}^* S_{mn}^{[2]} b_n^f = 0$. This is due to the fact that $\boldsymbol{a}^{\omega_f}$ fulfills the projection relation $\boldsymbol{a}^{\omega_f} = \boldsymbol{P}_f(\boldsymbol{a}^{\omega_f})$, which removes all $\boldsymbol{b}^f$ components from the linear response vector.

Finally, for the excited-state gradient

$$\frac{d\mathcal{C}}{dR_\beta} = \frac{\partial E_0}{\partial R_\beta} + b_m^{f*}\frac{\partial E_{mn}^{[2]}}{\partial R_\beta} b_n^f - \omega_f b_m^{f*}\frac{\partial S_{mn}^{[2]}}{\partial R_\beta} b_n^f - \bar{X}_m^* \frac{\partial E_m^{[1]}}{\partial R_\beta} \qquad (68)$$

**2.7. Implementation.** At variance with respect to what is required in standard linear and quadratic response calculations,[34,41,42] we need to implement the right-hand-side matrices $\boldsymbol{\eta}$ for the (adjoint) Lagrangian multiplier equations, as well as the final gradient expressions.

Comparing the expressions in eqs 61, 62, and 63, we write—in a somewhat self-explanatory notation—

$$\boldsymbol{\eta} = \boldsymbol{\eta}^{E3} + \boldsymbol{\eta}^{S3} + \boldsymbol{\eta}^{A2} + \boldsymbol{\eta}^{B2} \qquad (69)$$

where it is understood that the contribution $\boldsymbol{\eta}^{B2} = \mathbf{0}$ in the transition moment case, and $\boldsymbol{\eta}^{A2} = \boldsymbol{\eta}^{B2} = \mathbf{0}$ for the excited state.

The explicit expressions for the four contributions are found as

$$\boldsymbol{\eta}^{B2} = [[\mathbf{S}, \mathbf{B}^\dagger]_\mathrm{a}, \mathbf{S}]_\mathrm{D} \qquad (70)$$

$$\boldsymbol{\eta}^{A2} = -[[\mathbf{S}, \mathbf{A}]_\mathrm{b}, \mathbf{S}]_\mathrm{D} \qquad (71)$$

$$\boldsymbol{\eta}^{S3} = \mathbf{S}[\mathbf{D}, [\mathbf{a}, \mathbf{b}^\dagger]_\mathrm{S}]_\mathrm{S}\mathbf{S} = \mathbf{0} \qquad (72)$$

$$\boldsymbol{\eta}^{E3} = \mathbf{CDS} - \mathbf{SDC} \qquad (73)$$

with

$$\mathbf{C} = [\mathbf{S}, [\mathbf{F}, \mathbf{S}]_\mathrm{a}]_{\mathrm{b}^\dagger} + [\mathbf{S}, \mathbf{G}([\mathbf{a}, \mathbf{D}]_\mathrm{S})]_{\mathrm{b}^\dagger} + [\mathbf{S}, \mathbf{G}([\mathbf{b}^\dagger, \mathbf{D}]_\mathrm{S})]_\mathrm{a} + \mathbf{G}([[\mathbf{b}^\dagger, \mathbf{D}]_\mathrm{S}, \mathbf{a}]_\mathrm{S}) + \mathbf{T}^{\mathrm{xc}}([\mathbf{b}^\dagger, \mathbf{D}]_\mathrm{S}, [\mathbf{a}, \mathbf{D}]_\mathrm{S}) \qquad (74)$$

Above, **a** and **b** indicate, respectively, the linear response matrices for the $A$ and $B$ operators in the case of the linear response function; the *projected* linear response matrix at $\omega = \omega_f$, $\mathbf{a}^{\omega_f}$, and the eigenvector matrix $\mathbf{b}^f$ for the transition moment; and the excitation vector and its adjoint for the excited state gradient. Note that all $\boldsymbol{S}^{[3]}$ contributions actually vanish, as shown in Appendix A. This also applies for the additional contribution to the right-hand side originating from the last term in eq 62 in the case of the transition moment derivative

$$\bar{\omega}\mathbf{S}[\mathbf{D}, [\mathbf{b}^f, \mathbf{b}^{f\dagger}]_\mathrm{S}]_\mathrm{S}\mathbf{S} = \mathbf{0} \qquad (75)$$

For the final computational expressions of the property gradients, we need, in addition to the undifferentiated density, Fock and overlap matrices at the expansion point $\mathbf{R}_0$ and the integral matrices **A** and **B**, the differentiated one- and two-electron integral matrices $\mathbf{h}^{R_\beta}$ and $\mathbf{G}^{R_\beta}$, the differentiated overlap matrices $\mathbf{S}^{R_\beta}$, and the differentiated integral matrices $\mathbf{A}^{R_\beta}$ and $\mathbf{B}^{R_\beta}$, since we are considering a perturbation-dependent basis set where each atomic orbital is centered on a specific atom and thus depends on the spatial coordinates of the nuclei. Explicit expressions for the matrix elements of $\mathbf{A}^{R_\beta}$, when $A$ is the dipole moment operator, can be found in ref 44.

In ref 40, the derivatives of the individual matrices in the AO formulation were considered. It was there shown that the first derivative of the density, $\mathbf{D}^{R_\beta}(\mathbf{X})$, is given by the first derivative of the reference density matrix, $\mathbf{D}^{R_\beta}$, which, from the idempotency condition for **D**, is found to be

$$\mathbf{D}^{R_\beta} = -\mathbf{D}\mathbf{S}^{R_\beta}\mathbf{D} \qquad (76)$$

Comparing the three Lagrangian expressions in eqs 37, 38, and 39, we can thus write (once again in a self-explanatory notation)

$$\mathscr{L}^{A1} = A_m^{[1]} b_m = -\mathrm{Tr}\, \mathbf{A}[\mathbf{b}, \mathbf{D}]_\mathrm{S} \qquad (77)$$

$$\mathscr{L}^{B1} = a_m^* B_m^{[1]} = \mathrm{Tr}\,\mathbf{B}[\mathbf{D},\mathbf{a}^\dagger]_\mathrm{S} \tag{78}$$

$$\mathscr{L}^{E1} = \bar{X}_m^* E_m^{[1]} = \mathrm{Tr}\,\mathbf{F}[\mathbf{D},\bar{\mathbf{X}}^\dagger]_\mathrm{S} \tag{79}$$

$$\mathscr{L}^{E2} = a_m^* E_{mn}^{[2]} b_n = \mathrm{Tr}\,\mathbf{F}[[\mathbf{b},\mathbf{D}]_\mathrm{S},\mathbf{a}^\dagger]_\mathrm{S} + \mathrm{Tr}\,\mathbf{G}([\mathbf{b},\mathbf{D}]_\mathrm{S})[\mathbf{D},\mathbf{a}^\dagger]_\mathrm{S} \tag{80}$$

$$\mathscr{L}^{S2} = a_m^* S_{mn}^{[2]} b_n = \mathrm{Tr}\,\mathbf{a}^\dagger\mathbf{S}[\mathbf{D},\mathbf{b}]_\mathrm{S}\mathbf{S} \equiv -\mathrm{Tr}\,\mathbf{a}^\dagger\mathbf{S}[\mathbf{b},\mathbf{D}]_\mathrm{S}\mathbf{S} \tag{81}$$

When the perturbed densities

$$\mathbf{D}^b = [\mathbf{b},\mathbf{D}]_\mathrm{S} \tag{82}$$

$$\mathbf{D}^{b,R_\beta} = [\mathbf{b},\mathbf{D}^{R_\beta}]_\mathrm{S} + [\mathbf{b},\mathbf{D}]_{\mathrm{S}^{R_\beta}} \tag{83}$$

are introduced, the geometric derivatives of the above individual contributions can be written

$$\frac{\partial\mathscr{L}^{A1}}{\partial R_\beta} = \mathrm{Tr}\{\mathbf{b}([\mathbf{A}^{R_\beta},\mathbf{S}]_\mathrm{D} + [\mathbf{A},\mathbf{S}^{R_\beta}]_\mathrm{D} + [\mathbf{A},\mathbf{S}]_{\mathrm{D}^{R_\beta}})\} \tag{84}$$

$$\frac{\partial\mathscr{L}^{B1}}{\partial R_\beta} = \mathrm{Tr}\{\mathbf{a}^\dagger([\mathbf{B}^{R_\beta},\mathbf{S}]_\mathrm{D} + [\mathbf{B},\mathbf{S}^{R_\beta}]_\mathrm{D} + [\mathbf{B},\mathbf{S}]_{\mathrm{D}^{R_\beta}})\} \tag{85}$$

$$\frac{\partial\mathscr{L}^{S2}}{\partial R_\beta} = -\mathrm{Tr}\{\mathbf{a}^\dagger(\mathbf{S}^{R_\beta}\mathbf{D}^b\mathbf{S} + \mathbf{S}\mathbf{D}^{b,R_\beta}\mathbf{S} + \mathbf{S}\mathbf{D}^b\mathbf{S}^{R_\beta})\} \tag{86}$$

$$\begin{aligned}\frac{\partial\mathscr{L}^{E1}}{\partial R_\beta} &= \mathrm{Tr}\{\bar{\mathbf{X}}^\dagger([\mathbf{F}^{R_\beta},\mathbf{S}]_\mathrm{D} + [\mathbf{F},\mathbf{S}^{R_\beta}]_\mathrm{D} + \\ &\quad [\mathbf{F},\mathbf{S}]_{\mathrm{D}^{R_\beta}})\} + \mathrm{Tr}\{\mathbf{G}(\mathbf{D}^{R_\beta})[\mathbf{D},\bar{\mathbf{X}}^\dagger]_\mathrm{S}\} = \\ &\quad \mathrm{Tr}\{\bar{\mathbf{X}}^\dagger([\mathbf{F}^{R_\beta},\mathbf{S}]_\mathrm{D} + [\mathbf{F},\mathbf{S}^{R_\beta}]_\mathrm{D} + \\ &\quad [\mathbf{F},\mathbf{S}]_{\mathrm{D}^{R_\beta}})\} + \mathrm{Tr}\{\mathbf{G}([\mathbf{D},\bar{\mathbf{X}}^\dagger]_\mathrm{S})\mathbf{D}^{R_\beta}\}\end{aligned} \tag{87}$$

$$\begin{aligned}\frac{\partial\mathscr{L}^{E2}}{\partial R_\beta} &= \mathrm{Tr}\{\mathbf{F}^{R_\beta}[\mathbf{D}^b,\mathbf{a}^\dagger]_\mathrm{S} + \mathbf{F}[\mathbf{D}^{b,R_\beta},\mathbf{a}^\dagger]_\mathrm{S} + \mathbf{F}[\mathbf{D}^b,\mathbf{a}^\dagger]_{\mathrm{S}^{R_\beta}} + \\ &\quad \mathbf{G}(\mathbf{D}^{R_\beta})[\mathbf{D}^b,\mathbf{a}^\dagger]_\mathrm{S}\} + \mathrm{Tr}\{\mathbf{G}^{R_\beta}(\mathbf{D}^b)[\mathbf{D},\mathbf{a}^\dagger]_\mathrm{S} + \\ &\quad \mathbf{G}(\mathbf{D}^{b,R_\beta})[\mathbf{D},\mathbf{a}^\dagger]_\mathrm{S} + \mathbf{G}(\mathbf{D}^b)[\mathbf{D}^{R_\beta},\mathbf{a}^\dagger]_\mathrm{S} + \mathbf{G}(\mathbf{D}^b)[\mathbf{D},\mathbf{a}^\dagger]_{\mathrm{S}^{R_\beta}}\} = \\ &\quad \mathrm{Tr}\{\mathbf{F}^{R_\beta}[\mathbf{D}^b,\mathbf{a}^\dagger]_\mathrm{S} + \mathbf{F}[\mathbf{D}^{b,R_\beta},\mathbf{a}^\dagger]_\mathrm{S} + \mathbf{F}[\mathbf{D}^b,\mathbf{a}^\dagger]_{\mathrm{S}^{R_\beta}} + \\ &\quad \mathbf{G}([\mathbf{D}^b,\mathbf{a}^\dagger]_\mathrm{S})\mathbf{D}^{R_\beta}\} + \mathrm{Tr}\{\mathbf{G}^{R_\beta}(\mathbf{D}^b)[\mathbf{D},\mathbf{a}^\dagger]_\mathrm{S} + \\ &\quad \mathbf{G}([\mathbf{D},\mathbf{a}^\dagger]_\mathrm{S})\mathbf{D}^{b,R_\beta} + \mathbf{G}(\mathbf{D}^b)[\mathbf{D}^{R_\beta},\mathbf{a}^\dagger]_\mathrm{S} + \mathbf{G}(\mathbf{D}^b)[\mathbf{D},\mathbf{a}^\dagger]_{\mathrm{S}^{R_\beta}}\}\end{aligned} \tag{88}$$

where we take advantage of the fact that $\mathrm{Tr}\{\mathbf{G}(\mathbf{M})\mathbf{N}\} = \mathrm{Tr}\{\mathbf{G}(\mathbf{N})\mathbf{M}\}$ (and similarly for the differentiated $\mathrm{Tr}\{\mathbf{G}^{R_\beta}(\mathbf{M})\mathbf{N}\}$) to avoid computing two-electron Fock matrices for $R$-perturbed matrices, since it would require $3N$ Fock matrices computations.

Introducing the derivatives 84−88 into the generalized property expressions, eqs 65, 66, and 68, the explicit computational expressions for the property gradients are obtained, with the appropriate linear response matrices and eigenvector matrices taking the place of the generic $\mathbf{a}^\dagger$ and $\mathbf{b}$ matrices. For instance, the final computational expression for the linear-response function gradient reads

$$\begin{aligned}\frac{d\mathscr{C}}{dR_\beta} &= -\mathrm{Tr}\{\mathbf{b}([\mathbf{A}^{R_\beta},\mathbf{S}]_\mathrm{D} + [\mathbf{A},\mathbf{S}^{R_\beta}]_\mathrm{D} + [\mathbf{A},\mathbf{S}]_{\mathrm{D}^{R_\beta}})\} - \\ &\quad \mathrm{Tr}\{\mathbf{a}^\dagger([\mathbf{B}^{R_\beta},\mathbf{S}]_\mathrm{D} + [\mathbf{B},\mathbf{S}^{R_\beta}]_\mathrm{D} + [\mathbf{B},\mathbf{S}]_{\mathrm{D}^{R_\beta}})\} + \\ &\quad \mathrm{Tr}\{\mathbf{F}^{R_\beta}[\mathbf{D}^b,\mathbf{a}^\dagger]_\mathrm{S} + \mathbf{F}[\mathbf{D}^{b,R_\beta},\mathbf{a}^\dagger]_\mathrm{S} + \mathbf{F}[\mathbf{D}^b,\mathbf{a}^\dagger]_{\mathbf{S}^{R_\beta}} + \\ &\quad \mathbf{G}([\mathbf{D}^b,\mathbf{a}^\dagger]_\mathrm{S})\mathbf{D}^{R_\beta}\} + \mathrm{Tr}\{\mathbf{G}^{R_\beta}(\mathbf{D}^b)[\mathbf{D},\mathbf{a}^\dagger]_\mathrm{S} + \\ &\quad \mathbf{G}([\mathbf{D},\mathbf{a}^\dagger]_\mathrm{S})\mathbf{D}^{b,R_\beta} + \mathbf{G}(\mathbf{D}^b)([\mathbf{D}^{R_\beta},\mathbf{a}^\dagger]_\mathrm{S} + [\mathbf{D},\mathbf{a}^\dagger]_{\mathrm{S}^{R_\beta}})\} + \\ &\quad \omega\,\mathrm{Tr}\{\mathbf{a}^\dagger(\mathbf{S}^{R_\beta}\mathbf{D}^b\mathbf{S} + \mathbf{S}\mathbf{D}^{b,R_\beta}\mathbf{S} + \mathbf{S}\mathbf{D}^b\mathbf{S}^{R_\beta})\} - \\ &\quad \mathrm{Tr}\{\bar{\mathbf{X}}^\dagger\{[\mathbf{F}^{R_\beta},\mathbf{S}]_\mathrm{D} + [\mathbf{F},\mathbf{S}^{R_\beta}]_\mathrm{D} + [\mathbf{F},\mathbf{S}]_{\mathbf{D}^{R_\beta}}\}\} - \\ &\quad \mathrm{Tr}\{\mathbf{G}([\mathbf{D},\bar{\mathbf{X}}^\dagger]_\mathrm{S})\mathbf{D}^{R_\beta}\}\end{aligned} \tag{89}$$

with $\mathbf{a}^\dagger = \mathbf{a}^{\omega\dagger}$ and $\mathbf{b} = \mathbf{b}^\omega$.

The expression for the transition moment gradient is straightforwardly obtained from eq 89 by removing all terms involving the property integral matrix $\mathbf{B}$ and its geometric derivative $\mathbf{B}^R$, adding the $(\partial\mathscr{L}^{S2})/(\partial R_\beta)$ term multiplied by $\bar{\omega}$, with

$$\bar{\omega} = \frac{1}{2}\mathrm{Tr}\{\mathbf{A}[\mathbf{D},\mathbf{b}^f]_\mathrm{S}\} \tag{90}$$

and using $\omega_f$ instead of $\omega$ as factor in the third-last term. In this case, $\mathbf{a}^\dagger = \mathbf{a}^{\omega_f\dagger}$ and $\mathbf{b} = \mathbf{b}^f$.

In the excited state gradient, the terms involving the integral matrix $\mathbf{A}$ and its geometric derivative $\mathbf{A}^{R_\beta}$ also disappear and are replaced by the computational expression of the ground-state energy gradient[40]

$$\frac{\partial E_0}{\partial R_\beta} = \mathrm{Tr}\,\mathbf{D}\mathbf{h}^{R_\beta} + \frac{1}{2}\mathrm{Tr}\,\mathbf{D}\mathbf{G}^{R_\beta}(\mathbf{D}) + \mathrm{Tr}\,\mathbf{D}^{R_\beta}\mathbf{F} + h_\mathrm{nuc}^{R_\beta} \tag{91}$$

again with $\omega_f$ as the factor on the analog of the third-last term of eq 89, and with $\mathbf{a}^\dagger = \mathbf{b}^{f,\dagger}$ and $\mathbf{b} = \mathbf{b}^f$.

Note also that in Kohn−Sham theory the exchange-correlation contribution to the matrix $\mathbf{G}^{R_\beta}(\mathbf{D}^b)$ has a rather complicated expression, which is explicitly given in Appendix B.5.

If we replace the differentiated Hamilton operator (matrices) with a generic one-electron operator, $C$, in the above-given gradients, it is easy to prove that the computational expressions for the standard quadratic response function $\langle\langle A; B, C\rangle\rangle_{\omega,0}$ and its single and double residues are obtained. The proof is given in detail in section 2.4.2 of ref 28, starting from the expressions for the linear response function and transition moment magnetic gradients on LAOs, which, as mentioned in the Introduction, bear strong similarities with the gradients here considered.

As an addition to the discussion in ref 28, we report here the computational expression of a generic excited-state first-order property (component)—like for instance the excited state molecular electric dipole moment—as straightforwardly obtained from the excited-state gradient computational expression:

$$C_\alpha^f = \langle f|C_\alpha|f\rangle = \mathrm{Tr}\,\mathbf{D}\mathbf{C}_\alpha + C_{\alpha,\mathrm{nuc}} - \mathrm{Tr}\{[[\mathbf{b}^f,\mathbf{D}]_\mathrm{S},\mathbf{b}^{f\dagger}]_\mathrm{S} - [\mathbf{D},\bar{\mathbf{X}}^\dagger]_\mathrm{S}\}\mathbf{C}_\alpha \tag{92}$$

The first two terms in eq 92 correspond to the ground-state first-order property. It is well-known that excited-state first-order properties can also be computed as double residues of the quadratic response function.[42]

Also, the geometric gradient of any ground-state first-order property can be immediately obtained from the ground-state energy gradient in eq 91,

$$\frac{\partial C_\alpha}{\partial R_\beta} = \mathrm{Tr}\,\mathbf{D}\mathbf{C}_\alpha^{R_\beta} + \mathrm{Tr}\,\mathbf{D}^{R_\beta}\mathbf{C}_\alpha + C_{\alpha,\mathrm{nuc}}^{R_\beta} \tag{93}$$

## 3. Illustrative Results

The calculation of the geometric gradients of the linear response function, of its poles, and of its residues has been implemented within the linear-scaling development version of the Dalton code[34,45,46] at the Hartree−Fock and DFT levels of theory. The approach is general and encompasses the geometric gradient of any molecular property that can be related to the linear response function and its residues, like, for instance, the electric-dipole polarizability $\alpha_{\alpha\beta}(-\omega;\omega)$, whose geometric gradient is a key ingredient in the computational simulation of spectroscopic effects like Raman scattering[10] and coherent anti-Stokes Raman scattering.[11]

Various implementations of the polarizability as well as of the excited-state gradient have appeared in recent years, see for example refs 9, 11, and 23, whereas no analytic implementation of the electronic transition dipole moment derivative in a DFT framework has been presented. Note, however, an earlier CASSCF-based implementation of analytic derivatives of $\boldsymbol{\mu}_{kl}$ reported in ref 39. For this reason, as a specific illustrative application of the implementation, we will present and discuss here the results of a hybrid-functional DFT study of the vibronic fine structure of the $\tilde{\mathrm{X}}(^1\mathrm{A}_{1g}) - \tilde{\mathrm{A}}(^1\mathrm{B}_{2u})$ transition in the absorption spectrum of benzene. In the electric dipole approximation, this transition is Franck−Condon-forbidden and hence basically determined by the (first-order) Herzberg−Teller integrals and electronic transition dipole-moment derivatives. For such a study, the gradients of both the poles and residues have been used.

**3.1. Herzberg−Teller Contribution to One-Photon (UV) Spectra.** Quantum mechanical rovibronic transition moments, to which approximations are numerically calculated herein, are directly related to the experimentally determined integrated line strength of a rovibronic transition. Following ref 47, the integrated net absorption cross-section $G_{\mathrm{net}}$ reads

$$G_{\mathrm{net}} = \int_{\tilde{\nu}_1}^{\tilde{\nu}_2} \sigma_{\mathrm{net}}(\tilde{\nu})\tilde{\nu}^{-1}\,\mathrm{d}\tilde{\nu} = \int_{\tilde{\nu}_1}^{\tilde{\nu}_2} \sigma_{\mathrm{net}}(\tilde{\nu})\,\mathrm{d}\ln\tilde{\nu} \tag{94}$$

and thus depends on the wavenumber $\tilde{\nu}$-dependent absorption cross-section $\sigma_{\mathrm{net}}(\tilde{\nu})$, which is logarithmically integrated (see, e.g., ref 48 for a discussion) over a suitably chosen wavenumber interval that includes the entire absorption band.[47] The net absorption cross-section $\sigma_{\mathrm{net}}(\tilde{\nu})$ of a one-photon transition is connected via the Lambert−Beer law to the ratio between transmitted and incident spectral intensity $I_{\mathrm{tr},\tilde{\nu}}(\tilde{\nu})$ and $I_{0,\tilde{\nu}}(\tilde{\nu})$, respectively, (see ref 47 for a rigorous definition of these terms)

$$\sigma_{\mathrm{net}}(\tilde{\nu}) = \frac{1}{N_{\mathrm{A}}cl}\ln\left[\frac{I_{\mathrm{tr},\tilde{\nu}}(\tilde{\nu})}{I_{0,\tilde{\nu}}(\tilde{\nu})}\right] = \frac{A_{\mathrm{e}}(\tilde{\nu})}{N_{\mathrm{A}}cl} = \frac{\varepsilon(\tilde{\nu})\ln(10)}{N_{\mathrm{A}}} \tag{95}$$

where $N_{\mathrm{A}}$ is the Avogadro number, $c$ the amount of substance concentration of the absorbing species, $l$ the path length through the absorbing material, $A_{\mathrm{e}}$ the (Naperian) absorbance, and $\epsilon(\tilde{\nu})$ the molar (decadic) absorption coefficient. When stimulated emission can be neglected, the net integrated absorption cross-section between two energy levels is composed of the line strengths of the underlying individual transition processes between states $i$ and $j$, which are summed over and weighted according to the (fractional) population $p_i$ of the corresponding initial state:[47]

$$G_{\mathrm{net}} = \sum_{i,j} p_i G_{ij} \tag{96}$$

The individual integrated absorption cross-section is, via $G_{ij} = hB_{\tilde{\nu},ij}$, related to the Einstein coefficient $B_{\tilde{\nu},ij}$ for absorption and thus directly connected to the quantum mechanical transition moment. For the electric dipole transitions considered herein, the relation between electric transition dipole moment $\boldsymbol{M}_{ij}$ and integrated band strength $G_{ij}$ is[47]

$$G_{ij} = \frac{8\pi^3}{(4\pi\varepsilon_0)3hc_0}|\boldsymbol{M}_{ij}|^2 \tag{97}$$

with $|\boldsymbol{M}_{ij}|^2 = \sum_\alpha |\langle i|\hat{\mu}_\alpha|j\rangle|^2$ and $\alpha = x,\ y,\ z$. In contrast to the oscillator strength $f$, which is frequently used in UV/vis absorption spectroscopy, the integrated absorption band strength does not explicitly depend on the transition wavenumber. When $f_{ij}$ is defined as[47]

$$f_{ij} \approx \frac{m_{\mathrm{e}}c_0 8\pi^2\tilde{\nu}_{ij}}{e^2 3h}|\boldsymbol{M}_{ij}|^2 = \frac{(4\pi\varepsilon_0)m_{\mathrm{e}}c_0^2\tilde{\nu}_{ij}}{\pi e^2}G_{ij} \tag{98}$$

one obtains the following approximate relationship between oscillator strength and integrated band strength, which has been employed herein to convert previously reported values for $f_{ij}$ (or $f$) to integrated band strengths $G_{ij}$ (or $G_{\mathrm{net}}$):

$$f_{ij} \approx 1.1295835 \times 10^{-8}(\tilde{\nu}_{ij}/\mathrm{cm}^{-1})(G_{ij}/\mathrm{pm}^2);$$
$$f \approx 1.1295835 \times 10^{-8}(\tilde{\nu}_0/\mathrm{cm}^{-1})(G_{\mathrm{net}}/\mathrm{pm}^2) \tag{99}$$

with $\tilde{\nu}_0$ denoting the transition wavenumber of the corresponding band center.

The electric transition dipole moment $\boldsymbol{M}_{\kappa\lambda}$ between two rovibronic states characterized by the rovibronic wave functions $\Psi_\kappa(\mathbf{r},\mathbf{R})$ and $\Psi_\lambda(\mathbf{r},\mathbf{R})$, which depend on the collective electronic spatial coordinates $\mathbf{r}$ and the collective spatial coordinates $\mathbf{R}$ of the nuclei, is, in the adiabatic approximation, given by

$$\boldsymbol{M}_{\kappa\lambda} = \langle\kappa|\boldsymbol{\mu}|\lambda\rangle \approx \langle\kappa_k|\langle k|\boldsymbol{\mu}|l\rangle|\lambda_l\rangle = \langle\kappa_k|\boldsymbol{\mu}_{kl}|\lambda_l\rangle = \boldsymbol{M}_{\kappa k\lambda l} \tag{100}$$

with the adiabatic electronic wave functions $\psi_k(\mathbf{r};\mathbf{R})$ and $\psi_l(\mathbf{r};\mathbf{R})$ depending explicitly on the spatial coordinates of the electrons and parametrically on the spatial coordinates of the nuclei. The wave functions $\chi_{\kappa,k}(\mathbf{R})$ and $\chi_{\lambda,l}(\mathbf{R})$ for the motion of the nuclei depend, like the electronic transition dipole moment $\mu_{kl}(\mathbf{R})$, only on the coordinates of the various nuclei.

If a Taylor series expansion is applied to $\boldsymbol{\mu}_{kl}(\mathbf{R})$, for instance around the equilibrium molecular structure $\mathbf{R}_0$ of the initial electronic state, the expansion yields

$$\boldsymbol{\mu}_{kl}(\mathbf{R}) = \boldsymbol{\mu}_{kl}(\mathbf{R}_0) + \sum_{\beta} \frac{\partial \boldsymbol{\mu}_{kl}(\mathbf{R})}{\partial R_{\beta}}\bigg|_{\mathbf{R}=\mathbf{R}_0} (R_{\beta} - R_{0,\beta}) + ... \tag{101}$$

By inserting this expansion into eq 100, the electric transition dipole moment is expressed as a sum of Franck−Condon and Herzberg−Teller contributions corresponding to the terms involving the electronic transition dipole moment and the first derivative of the electronic transition dipole moment with respect to displacements of the nuclei computed at the molecular equilibrium structure, respectively, in addition to higher-order terms. If the latter are neglected, one obtains

$$\boldsymbol{M}_{\kappa k \lambda l} = \boldsymbol{\mu}_{kl}(\mathbf{R}_0)\langle \kappa_k | \lambda_l \rangle + \sum_{\beta} \frac{\partial \boldsymbol{\mu}_{kl}(\mathbf{R})}{\partial R_{\beta}}\bigg|_{\mathbf{R}=\mathbf{R}_0} \langle \kappa_k | (\hat{R}_{\beta} - \hat{R}_{0,\beta}) | \lambda_l \rangle \tag{102}$$

Within the Born−Oppenheimer adiabatic approximation, which is employed in this work, the transition dipole moment and its first derivatives with respect to nuclear displacements can be computed analytically using the DFT-based framework developed in this paper. We note in passing that, in addition to the Herzberg−Teller terms, terms arising through diabatic (frequently called nonadiabatic) coupling also contribute to the transition dipole moment[49] to this order, which are, however, neglected in the present study.

Franck−Condon integrals $\langle \kappa_k | \lambda_l \rangle$ and Herzberg−Teller integrals $\langle \kappa_k | (\hat{R}_{\beta} - \hat{R}_{0,\beta}) | \lambda_l \rangle$ involve the wave functions of the motions of the nuclei. If the vibrational motion is assumed to be harmonic and separable from the rotational and translational motion, the states $|\kappa_k\rangle$ and $|\lambda_l\rangle$ are expressed as direct products of multidimensional harmonic oscillator states $|\boldsymbol{v}\rangle$ and $|\boldsymbol{v}'\rangle$ (with quantum numbers $v_i$ and $v_i'$ for the various harmonic oscillators in the initial and final states, respectively) and corresponding rotational-translational states. The latter are not explicitly considered herein. To this end, the spatial coordinates of the nuclei $\mathbf{R}$ are replaced by the vibrational mass-weighted normal coordinates, denoted as $\mathbf{Q}$ and $\mathbf{Q}'$, in which the harmonic vibrational force fields of the two electronic states involved are diagonal, as well as by the Euler angles for rotation and by the spatial coordinates of the center of mass. Recently, a coherent state-based generating function approach for efficiently computing vibronic transition profiles within the Franck−Condon and Herzberg−Teller approximation (and beyond) has been outlined for Duschinsky rotated multidimensional harmonic oscillators at finite temperatures and at 0 K.[50] We employ this coherent state-based generating function approach implemented in the vibronic structure program hotFCHT,[13,51,52] which offers both a time-independent and time-dependent route to electric transition properties. The vibronic profile generating function $G_{\mathrm{FCHT}}(\mathbf{Z}; \boldsymbol{\Lambda})$, not to be confused with the integrated band strengths $G_{ij}$, $G_{\mathrm{net}}$, and $G_{\mathrm{total}}$, consists of three parts: one containing the Franck−Condon factor, the second one the Franck−Condon/Herzberg−Teller integrals, and the last one the (first-order) Herzberg−Teller term,

$$\begin{aligned} G_{\mathrm{FCHT}}(\mathbf{Z};\boldsymbol{\Lambda}) = {} & |\boldsymbol{\mu}_{kl}(\mathbf{Q}_0)|^2 G_{\mathrm{FC}}^{K}(\mathbf{Z};\boldsymbol{\Lambda})^{(\hat{1},\hat{1})} + \\ & 2\sum_{\beta} \boldsymbol{\mu}_{kl}(\mathbf{Q}_0)\cdot\left(\frac{\partial \boldsymbol{\mu}_{kl}}{\partial Q_{\beta}}\right)_{\mathbf{Q}=\mathbf{Q}_0} G_{\mathrm{FC/HT}}^{K}(\mathbf{Z};\boldsymbol{\Lambda})^{(\hat{Q}_{\beta},\hat{1})} + \\ & \sum_{\beta,\gamma}\left(\frac{\partial \boldsymbol{\mu}_{kl}}{\partial Q_{\beta}}\right)_{\mathbf{Q}=\mathbf{Q}_0}\cdot\left(\frac{\partial \boldsymbol{\mu}_{kl}}{\partial Q_{\gamma}}\right)_{\mathbf{Q}=\mathbf{Q}_0} G_{\mathrm{HT}}^{K}(\mathbf{Z};\boldsymbol{\Lambda})^{(\hat{Q}_{\beta},\hat{Q}_{\gamma})} \end{aligned} \tag{103}$$

where $\mathbf{Z}$ contains the generating function parameters, $\boldsymbol{\Lambda}$ is related to a thermal integration kernel $K$, $\hat{1}$ is the identity operator, and $\hat{Q}_{\beta}$ is the position operator corresponding to the $\beta$th normal coordinate $Q_{\beta}$. Details of the approach and the definition of the various terms in eq 103 can be located in ref 50.

**3.2. Computational Details.** As a test case, we present the results of calculations of the vibrational fine structure in the $\tilde{\mathrm{X}}(^1\mathrm{A}_{1g}) - \tilde{\mathrm{A}}(^1\mathrm{B}_{2u})$ one-photon UV absorption spectrum of benzene at 0 K. This transition is Franck−Condon-forbidden in the electric transition dipole approximation, that is, $\boldsymbol{\mu}_{kl}(\mathbf{Q_0}) = \mathbf{0}$ at the $D_{6h}$ symmetric equilibrium structure of the initial and final electronic states, and becomes allowed due to Herzberg−Teller vibronic coupling. The time-dependent Hartree−Fock and the time-dependent density functional theory methods, the latter using the B3LYP functional[53,54] as well as its Coulomb attenuated variant camB3LYP,[55] were exploited for the electronic structure calculation within the linear-scaling development version of the Dalton program. As a basis set, the triple-$\zeta$ valence basis set with polarization functions (TZVP) of ref 56 was used. The grid employed is based on the original Becke partitioning and the radial grid of ref 57 multiplied by an angular Lebedev grid.[58−60] The grid is pruned for small $R$ in order to avoid too many grid points with small weights, and the radial integration threshold was chosen to be $10^{-13}$. The angular expansion order was chosen to be 42.

Equilibrium structures in the electronic ground state were obtained using analytic derivatives of the total electronic energy with respect to nuclear displacements using a convergence threshold of $10^{-5}$ $E_{\mathrm{h}}\, a_0^{-1}$ for the norm of the gradient and $10^{-4}$ $E_{\mathrm{h}}\, a_0^{-1}$ for its largest component. The total energy of each cycle was optimized to $10^{-6}$ $E_{\mathrm{h}}$. Harmonic force fields of the electronic ground state were calculated using analytic second derivatives with thresholds of $10^{-7}$ when solving the linear response equations. The equilibrium structures in the electronically excited state were computed using analytical derivatives of the excited-state energy with respect to displacements of the nuclei (see eq 68). The excitation energies were converged until changes remained below $10^{-6}$ $E_{\mathrm{h}}$, and the norm of the final excited state gradient was below $10^{-5}$ $E_{\mathrm{h}}\, a_0^{-1}$. The harmonic force constant matrix of the electronically excited states was computed using central numerical derivatives of analytic gradients of the electronic energy with respect to displacements of the nuclei. The finite size of the corresponding displacements was chosen as 0.01 $a_0$. The masses employed were those of the

***Table 1.*** Harmonic Vibrational Wavenumbers (in cm$^{-1}$) for $e_{2g}$ Modes of the $^1A_{1g}$ and $^1B_{2u}$ States of Benzene[a]

| state | mode | B3LYP/ TZVP | camB3LYP/ TZVP | HF/ TZVP | CASSCF/ DZP[b] | exptl[c] |
|---|---|---|---|---|---|---|
| $^1A_{1g}$ | $\nu_6$ | 625 | 631/632 | 666 | 646 | 608 |
| | $\nu_9$ | 1202 | 1212 | 1284 | 1263 | 1178 |
| | $\nu_8$ | 1634 | 1673/1674 | 1771 | 1730 | 1600 |
| | $\nu_7$ | 3169 | 3196 | 3328 | 3369 | 3057 |
| $^1B_{2u}$ | $\nu_6'$ | 533/539 | 538/539 | 563 | 575 | 521 |
| | $\nu_9'$ | 1181/1186 | 1192/1193 | 1265 | 1237 | 1148 |
| | $\nu_8'$ | 1565/1566 | 1606 | 1712 | 1665 | 1516 |
| | $\nu_7'$ | 3192/3194 | 3220 | 3356 | 3389 | 3077 |

[a] A pair of numbers separated by a slash is given when the vibrational wavenumbers for a pair of normal modes that are supposed to be degenerate are different due to numerical reasons (in the current work, we do not fully exploit point group symmetry). [b] Ref 13. [c] Taken from the compilation of data reported in ref 69.

most abundant isotopes ($^{12}$C,$^1$H). Electronic transition dipole moments (which vanish in the present case) and their analytic derivatives with respect to displacements of the nuclei were computed according to eq 66 at the computed equilibrium structures of the initial state. Symmetry has not been used when computing the electronic transition dipole moment derivatives, as well as the energy gradient and the harmonic force field in the electronically excited state. The other calculations were performed by taking advantage of the Abelian $D_{2h}$ point group symmetry.

The plotted spectral profiles of the one-photon absorption spectra were determined within the more efficient time-dependent approach. In the evaluation of the Fourier transformation of the Lorentzian weighted time-correlation function (TCF), corresponding to the Lorentzian weighted eq 103 in the time domain, the FFTW[61] library (version 3.1.2) was used for the fast Fourier transformation with a grid size of $2^{15}$, a time increment of $t \sim 1.0$ fs, and a time interval of ($-16.384$ ps, 16.384 ps). The real part of the Fourier transformed TCF was taken, and its norm was plotted after weighting with the transition wavenumber as the wavenumber-dependent absorption cross-section $\sigma(\tilde{\nu})$. Integrated absorption cross-sections of the individual vibronic transitions reported in the tables were directly computed within the time-independent framework.

**3.3. Discussion of the Results.** The computed harmonic vibrational wavenumbers of the modes which transform according to the irreducible representation $e_{2g}$ (corresponding to $\nu_6$, $\nu_7$, $\nu_8$, and $\nu_9$ in the nomenclature of Wilson[62] and $\nu_{18}$, $\nu_{15}$, $\nu_{16}$, and $\nu_{17}$ in the nomenclature of Herzberg, respectively) are reported in Table 1 for the electronic states $\tilde{X}(^1A_{1g})$ and $\tilde{A}(^1B_{2u})$. Only these doubly degenerate modes of benzene are capable of inducing intensity for this electric dipole transition via first-order Herzberg−Teller vibronic coupling, whereas in the second order, other modes such as $e_{1g}$, $e_{1u}$, and $e_{2u}$ can act as inducing modes (see, e.g., ref 63). For comparison, we also report the experimental fundamental wavenumbers and the harmonic wavenumbers computed in ref 13 in the complete active space-self-consistent field (CASSCF) framework for the four $e_{2g}$ modes.

The Herzberg−Teller absorption profiles computed for a temperature of 0 K on the basis of the results from the various electronic structure approaches are plotted in Figure 1. The

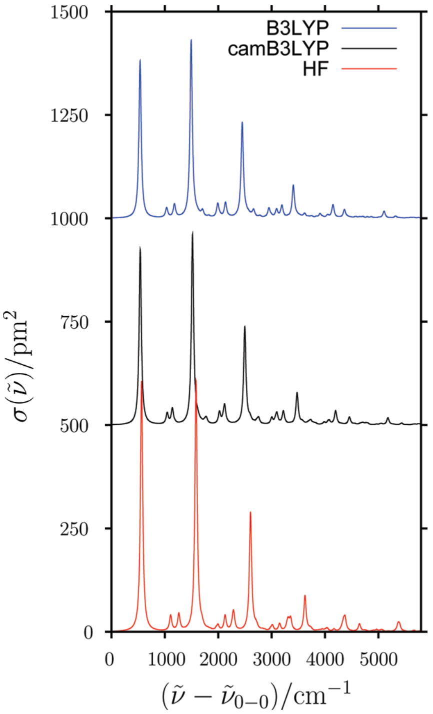


***Figure 1.*** Calculated absorption cross-sections $\sigma(\tilde{\nu})$ (in pm$^2$) as a function of the wavenumber (in cm$^{-1}$) in excess of the 0−0 transition wavenumber $\tilde{\nu}_{0-0}$ as obtained from TDDFT (using the B3LYP and the camB3LYP functional) and from TDHF. Cross-sections were computed using the coherent state generating function approach in the time-dependent picture by exploiting the time-correlation function. For the graphical representation, the cross-sections computed with the B3LYP and camB3LYP hybrid density functionals were shifted by an increment of 1000 pm$^2$ and 500 pm$^2$, respectively. A Lorentzian line shape function with full-width at half-maximum of 50 cm$^{-1}$ was employed, and the experimental value from ref 63 of $\tilde{\nu}_{0-0} = 38\,086$ cm$^{-1}$ was used in converting the HT profile to wavenumber-dependent absorption cross-sections, which corresponds for an isolated band after logarithmic integration over a suitable wavenumber interval to the integrated absorption cross-section *G* of the vibronic band. The values of *G* reported in the tables were computed, however, directly in the time-independent picture via eqs 102 and 97.

0−0 transition in the $\tilde{X}(^1A_{1g}) - \tilde{A}(^1B_{2u})$ absorption spectrum of benzene is Franck−Condon-forbidden in the electric dipole approximation. The first vibronic band around $\tilde{\nu}_{0-0} + 550$ cm$^{-1}$ serves as a so-called false origin brought about by the $6^1_0$ transition (using Wilson's mode enumeration), on which the most prominent progression $6^1_0 1^{n'}_0$ builds, extending (visibly) up to about 5500 cm$^{-1}$ above the 0−0 transition wavenumber. This progression is shortest (in terms of relative cross sections) at the Hartree−Fock level (abbreviated as HF below and in the tables) and longest for the B3LYP functional, which differs only slightly from the camB3LYP result. This finding is in line with the predicted change in the C−C bond length, for which HF gives only $\Delta r_{C-C} = 2.82$ pm, whereas B3LYP gives $\Delta r_{C-C} = 3.08$ pm and camB3LYP $\Delta r_{C-C} = 2.99$ pm. These C−C bond length elongations upon electronic excitation are significantly smaller than those predicted in ref 13 at the CASSCF level (3.8 pm), in ref.64 at the CCSD level (3.3 pm) and in ref 39 at the CASPT2 level (3.6 pm). This strongly impacts the overall shape of the progression in the breathing mode $\nu_1$, which is predicted to be too short as compared to the multireference and coupled cluster methods, as well as the experiment. Less prominent progressions in this mode are built on the other false origins $7^1_0$, $8^1_0$, and $9^1_0$ as well as on combination bands such as $6^1_0 16^2_0$ and $6^1_0 17^2_0$. For selected vibronic bands, we report in Table 2 the computed integrated absorption band strengths as well as their relative values and compare these to earlier computational results and experimental measurements. To facilitate this comparison, we have converted previously reported oscillator strengths to integrated (net) absorption band strengths using eq 99. When comparing experimental and theoretical band strengths, we do, however, not correct or account for temperature-dependent populations of the initial state, limited experimental resolutions, and other factors.

According to the Herzberg−Teller sum rule (see, e.g., ref 15), the total integrated absorption band strength $G_{\mathrm{total}}$ for the $\tilde{\mathrm{X}}(^1\mathrm{A}_{1g}) - \tilde{\mathrm{A}}(^1\mathrm{B}_{2u})$ transition of benzene at 0 K becomes in the harmonic approximation

$$G_{\mathrm{total}} = \frac{8\pi^3}{(4\pi\varepsilon_0)3hc_0} \sum_{\beta} \frac{\hbar}{2\omega_\beta} \left| \left( \frac{\partial \boldsymbol{\mu}_{kl}}{\partial Q_\beta} \right)_{\mathbf{Q}=\mathbf{Q}_0} \right|^2 \qquad (104)$$

where $\omega_\beta$ is the harmonic frequency corresponding to the normal mode coordinate $Q_\beta$ of the initial electronic state, and where the sum runs over all components of degenerate normal modes, including, for example, $\nu_{6a}$ and $\nu_{6b}$. Alternatively, one weights the various terms with the degeneracies of the normal modes and sums over the contributions from the modes to obtain $G_{\mathrm{total}}$. The total integrated cross-section can in this approximation be partitioned into contributions from each normal mode. Accordingly, the magnitude of the contribution of the various vibrational modes is determined by the harmonic frequency and the norm square of the first derivative of the electronic transition dipole moment. In Table 3, the contributions of the various $e_{2g}$ vibrational modes to the integrated band strength from eq 104 as well as their sum are presented.

From Tables 2 and 3 it is evident that the $\nu_6$ mode accounts for the major part of the first-order HT-induced absorption cross-section. One of the main differences between the

***Table 2.*** Computed and Experimental Integrated Absorption Band Strengths (in pm$^2$) for Selected Vibronic Transitions between the Electronic Singlet Ground State ($^1$A$_{1g}$) and the Lowest Excited Singlet State ($^1$B$_{2u}$) of Benzene[a]

| trans. | B3LYP/ TZVP | camB3LYP/ TZVP | HF/ TZVP | CASSCF/ DZP[b] | CASPT2/ ANO[c] | exp.[d] | exp.[e] | exp.[f] | exp.[g] | exp.[h] |
|---|---|---|---|---|---|---|---|---|---|---|
| $6^1_0$ | 0.784(1.00) | 0.863(1.00) | 1.232(1.00) | 0.110(1.00) | 0.205(1.00) | 0.314(1.00) | 0.20(1.00) | 0.17(1.00) | (1.00) | (1.000) |
| $6^1_0 1^1_0$ | 0.861(1.10) | 0.910(1.05) | 1.199(0.97) | 0.189(1.72) | 0.309(1.50) | 0.419(1.33) | 0.338(1.68) | 0.20(1.12) | (0.99) | (0.903) |
| $6^1_0 1^2_0$ | 0.448(0.57) | 0.455(0.53) | 0.554(0.45) | 0.156(1.41) | 0.220(1.07) | 0.293(0.93) | 0.287(1.42) | 0.14(0.75) | (0.94) | (0.889) |
| $6^1_0 1^3_0$ | 0.147(0.19) | 0.143(0.17) | 0.162(0.13) | 0.082(0.74) | 0.099(0.48) | 0.182(0.58) | 0.15(0.72) | 0.07(0.37) | (0.49) | (0.010) |
| $6^1_0 1^4_0$ | 0.034(0.04) | 0.032(0.04) | 0.033(0.03) | 0.031(0.28) | 0.031(0.15) | 0.119(0.38) | 0.07(0.36) | 0.02(0.11) | (0.13) | |
| $7^1_0$ | 0.052(0.067) | 0.058(0.067) | 0.055(0.044) | 0.006(0.054) | 0.011(0.05) | | | 0.01(0.07) | (0.06) | (0.034) |
| $8^1_0$ | 0.012(0.015) | 0.008(0.008) | 0.006(0.005) | 0.004(0.033) | 0.001(0.005) | | | | | (0.006) |
| $9^1_0$ | 0.005(0.007) | 0.003(0.004) | 0.0001(0.0001) | 0.003(0.028) | 0.008(0.04) | | | | (0.02) | (0.018) |

[a] Relative values for the integrated absorption cross-sections are reported in parentheses. [b] Values obtained using the same input data set as in ref 13. [c] Ref 39, oscillator strengths reported in ref 39 (for the CASPT2 equilibrium structures combined with CASSCF harmonic force fields and electronic transition dipole moments) were converted according to eq 99 using the reported computed transition wavenumbers. [d] Ref 67, oscillator strengths reported in ref 67 were converted according to eq 99 using the experimental transition wavelengths given in Table 1 of this reference. [e] Ref 68, oscillator strengths reported in ref 68 were converted according to eq 99 using the experimental transition wavelengths reported in Table 1 of ref 67. [f] Ref 66, relative intensities (oscillator strengths) reported in ref 66 were converted according to eq 99 using the experimental transition wavenumbers reported in Figure 1 of ref 66 for all transitions except for $7^1_0$, where the transition wavenumber of ref 65 was employed; integrated cross-sections were estimated from the data reported in ref 66 assuming an amount of substance concentration of $1.47 \times 10^{-3}$ mol l$^{-1}$ of benzene at the given temperature. [g] Following ref 63, the relative intensities reported in ref 70 were converted according to eq 99 using the experimentally derived transition wavenumbers ($\nu_{\mathrm{origin}}$) reported in Table 6 of ref 63. [h] Ref 65, the reported relative peak heights from fluorescence excitation spectra (according to those authors, only rough estimates of the intensities) were corrected for the wavenumber dependence using the transition wavenumbers reported in Table 2 of ref 65.

***Table 3.*** Contributions (in pm$^2$) of the Various $e_{2g}$ Vibrational Normal Modes to the Total Integrated Absorption Cross-Section $G_{\mathrm{total}}$ (calculated according to eq 104) for the ($^1$A$_{1g}$ → $^1$B$_{2u}$) Transition of Benzene at 0 K[a]

| mode | B3LYP/TZVP | camB3LYP/TZVP | HF/TZVP | CASSCF/DZP[b] | CASPT2/ANO[c] | exp.[d] | exp.[e] | exp.[f] |
|---|---|---|---|---|---|---|---|---|
| $\nu_6$ | 2.893[2.27] | 3.076[2.40] | 3.988[3.18] | 0.708[0.57] | [0.86] | [1.33] | [1.05] | [0.60] |
| $\nu_7$ | 0.202 | 0.216 | 0.185 | 0.039 | | | | [0.04] |
| $\nu_8$ | 0.052 | 0.033 | 0.018 | 0.026 | | | | |
| $\nu_9$ | 0.014 | 0.008 | 0.000 | 0.016 | | | | |
| sum | 3.160 | 3.334 | 4.191 | 0.789 | | | 3 | |

[a] Values in brackets correspond to the contribution from a limited number of members of an $a^1_0 1^{n'}_0$ progression ($a$ = 6, 7, 8, 9). [b] Values obtained using the same input data set as in ref 13. [c] Ref 39, the sum of integrated absorption cross-sections as given in the present Table 2 for the $6^1_0 1^{n'}_0$ progression with $n' = 0{-}4$. [d] Ref 67, the sum of integrated absorption cross-sections as given in the present Table 2 for the $6^1_0 1^{n'}_0$ progression with $n' = 0{-}$ 4; if $n' = 5$ of Table 1 in ref 67 is included ($G \approx 0.1$ pm$^2$), one obtains a values of (1.43 pm$^2$). [e] Ref 68, the sum of integrated absorption cross-sections as given in the present Table 2 for the $6^1_0 1^{n'}_0$ progression with $n' = 0{-}4$; the value for $G_{\mathrm{total}}$ was obtained by converting the oscillator strength reported in ref 68 according to eq 99 by employing the transition energy for $\varepsilon_{\max}$ given in Table 2 of that work. [f] Ref 66, the sum of integrated absorption cross-sections given in the present Table 2 for the $6^1_0 1^{n'}_0$ progression with $n' =$ 0−4; the partial contribution from the $7^1_0 1^{n'}_0$ progression with $n' = 0{-}2$ was obtained from the data of ref 66 as described in Table 2 above using the $7^1_0$ transition wavenumber of ref 65 and the wavenumber increments give in Table 2 of ref 66.

CASSCF results reported in ref 13 and experimental results[65] is the $8_0^1$ band, whose relative integrated band strength contribution was predicted to be too large when compared to the experimental intensity estimates obtained from peak height measurements in fluorescence excitation spectra of jet-cooled benzene.[65] The agreement seems to improve in the present work, at the price, however, of the $9_0^1$ bands now being significantly less intense (on relative terms) than the experimental estimate, and even the tendency ($7_0^1 > 8_0^1 > 9_0^1$) differs from that of the experiment ($7_0^1 > 9_0^1 > 8_0^1$). When the relative integrated absorption cross-sections of $7_0^1$ computed with the hybrid functionals are compared to experimental values[66] obtained from integration of the band profiles, the agreement seems almost perfect. The computed absolute values of the integrated cross-sections are, however, too large by more than a factor of 2 for the $6_0^1$ transition, when judged by the experimental data of refs 67 and 68. The progression in the C−C stretching mode $\nu_1$ built on the $6_0^1$ transition appears to be too short when compared to the experiment. While the integrated band strengths of the $6_0^1 1_0^1$ are found to be the largest of the $6_0^1 1_0^{n'}$ progression at the hybrid functional level, in agreement with the experiment reported in refs 66−68, the integrated absorption cross-section ratio for $6_0^1 1_0^2$/$6_0$ is found to be 1/2 only, whereas experimentally the ratio is 3/4 or even larger. This too-short progression is rooted in an underestimated C−C bond length elongation upon electronic excitation.

The intensity induced by mode $\nu_6$ via first-order Herzberg−Teller vibronic coupling appears overestimated (by about a factor of 2) according to the sum rule of eq 104. As the harmonic vibrational wavenumber of $\nu_6$ in the electronic ground state is in reasonable agreement with experimental results (too large by about 3% on the B3LYP level), the first derivative of the electronic transition dipole moments appears not to be well-estimated or just the shape of the normal modes may not be adequately described. Interestingly, the total integrated absorption cross-section of the $\tilde{X}$ ($^1A_{1g}$) − $\tilde{A}$($^1B_{2u}$) transition is obtained at the hybrid functional level in reasonable agreement with the value deduced from the oscillator strength reported in ref 68. If one assumes this agreement to be fortuitous, it could indicate that the inducing strength of the other modes is strongly underestimated at the DFT level (which appears, however, not to be the case) or it could point to a significant higher-order Herzberg−Teller vibronic coupling contribution and possibly also a diabatic coupling contribution that gives rise to additional intensity stealing, or it could hint to some pronounced intensity redistribution due to various resonances from progressions involving $\nu_6$ to other bands. Also, finite temperature effects, which have been neglected in the current study, and effects due to finite resolution, could play a role.

## 4. Conclusion

We have presented a derivation and implementation of third-order response properties that are connected to geometric derivatives of second-order response properties. The implementation is based on the exponential parametrization of the density matrix in the atomic orbital basis within time-dependent Hartree−Fock and density functional response theory. The formulation is linearly scaling for sufficiently sparse matrices.

We have demonstrated the applicability of the approach by considering the UV/vis absorption spectrum for the electronic ground state to the energetically lowest excited singlet state of benzene, a transition which is entirely dominated by vibronic coupling, here described by the first-order Herzberg−Teller corrections. The determination of the HT corrections requires the computation of the electronic excited state geometry and Hessian, as well as the geometric derivative of the electronic transition dipole moment. This is the first implementation of the analytic computation of first-order Herzberg−Teller corrections at the DFT level of theory, offering a straightforward way of overcoming the phase (and hence sign) uncertainties that can appear when determining the HT corrections by numerical derivative techniques, as well as avoiding the high computational cost, especially for systems with many degrees of freedom, of the numerical approach. Another advantage of our analytical (linear-scaling) implementation of the transition moment gradient is the possibility to combine it with a numerical derivative scheme, as already done for instance for the excited-state Hessian, to compute the second-order contributions (with respect to the nuclear coordinates) to the transition dipole. This would allow the investigation of vibronic effects beyond the linear Herzberg−Teller approximation, effects that are expected to be specifically relevant for highly symmetric systems, such as the one here investigated.

The results obtained for the Franck−Condon-forbidden electric dipole transition from the electronic ground state to the lowest excited singlet state of benzene are in qualitative agreement with experiment. Quantitatively, however, for the selected combinations of functionals (B3LYP, camB3LYP, HF) and basis set (TZVP), the total integrated absorption cross-section is found to be too large by about a factor of 2, and the predicted relative HT induction strength of the various $e_{2g}$ modes is also not fully satisfactory. As the variability-limited accuracy of time-dependent density functional theory in the prediction of UV/vis transition wavenumbers of conjugated $\pi$ systems is known, a thorough benchmark study on the performance of the various functionals in predicting quantitatively the first-order Herzberg−Teller contribution is required, for which the approaches and implementation presented in this work provide a valuable starting point.

**Acknowledgment.** This work has been supported by the Lundbeck Foundation and the Danish Research Council (Grant No. 9901973) and by the European Research and Training Network NANOQUANT ("Understanding Nano-Materials From the Quantum Perspective", Contract number MRTN-CT-2003-506842). S.C. acknowledges affiliation to the Instituto Nazionale Scienze dei Materiali (INSTM) and financial support from the Italian Ministero dell'Università Ricerca within the PRIN2006 (Programmi di ricerca di interesse nazionale) funding scheme. R.B. thanks the Volkswagen Foundation for financial support, acknowledges the Center for Scientific Computing (CSC) Frankfurt for computer time, and is indebted to Jürgen Stohner as well as Mark Thomson for discussions. K.R. has received support from the Norwegian Research Council through a Centre of Excellence Grant (Grant No. 179568/V30), a YFF grant (Grant No. 162746/V00), and a VIBRON Grant (Grant No. 177558/V30). This work has also received support from NordForsk (Grant No. 070253). We also thank Jeppe Olsen and Lea S. Thøgersen for their invaluable contributions at an early stage of this project.

## Appendix A. The Second-Order Renormalization Contribution

All $\boldsymbol{S}^{[3]}$ contributions to the right-hand sides in eqs 61, 62, and 63 vanish due to the fact that all of the involved vectors contain no redundant parameters, as they fulfill the projection relation 11,

$$\mathbf{a} = \mathscr{P}(\mathbf{a}) \equiv \mathbf{P}_{\mathrm{o}}\mathbf{a}\mathbf{P}_{\mathrm{v}}^{\mathrm{T}} + \mathbf{P}_{\mathrm{v}}\mathbf{a}\mathbf{P}_{\mathrm{o}}^{\mathrm{T}} \tag{105}$$

$$\mathbf{b}^{\mathrm{T}} \equiv \mathbf{P}_{\mathrm{v}}\mathbf{b}^{\mathrm{T}}\mathbf{P}_{\mathrm{o}}^{\mathrm{T}} + \mathbf{P}_{\mathrm{o}}\mathbf{b}^{\mathrm{T}}\mathbf{P}_{\mathrm{v}}^{\mathrm{T}} \tag{106}$$

The $\boldsymbol{\eta}^{S3}$ contribution was given in eq 72

$$\boldsymbol{\eta}^{S3} = \mathbf{S}[\mathbf{D}, [\mathbf{a}, \mathbf{b}^{\mathrm{T}}]_{\mathrm{S}}]_{\mathrm{S}}\mathbf{S} = \mathbf{S}(\mathbf{P}_{\mathrm{o}}[\mathbf{a}, \mathbf{b}^{\mathrm{T}}]_{\mathrm{S}} - [\mathbf{a}, \mathbf{b}^{\mathrm{T}}]_{\mathrm{S}}\mathbf{P}_{\mathrm{o}}^{\mathrm{T}})\mathbf{S} \tag{107}$$

Using the idempotency ($\mathbf{P}_{\mathrm{o}}^2 = \mathbf{P}_{\mathrm{o}}$ and $\mathbf{P}_{\mathrm{v}}^2 = \mathbf{P}_{\mathrm{v}}$) and orthogonality relations ($\mathbf{P}_{\mathrm{o}}\mathbf{P}_{\mathrm{v}} = \mathbf{P}_{\mathrm{v}}\mathbf{P}_{\mathrm{o}} = \mathbf{0}$ and $\mathbf{P}_{\mathrm{o}}^{\mathrm{T}}\mathbf{S}\mathbf{P}_{\mathrm{v}} = \mathbf{P}_{\mathrm{v}}^{\mathrm{T}}\mathbf{S}\mathbf{P}_{\mathrm{o}} = \mathbf{0}$), the contribution $\boldsymbol{\eta}^{S3}$ may be rewritten as

$$\begin{aligned}\boldsymbol{\eta}^{S3} = \mathbf{S}\{&(\mathbf{P}_{\mathrm{o}}\mathbf{a}\mathbf{P}_{\mathrm{v}}^{\mathrm{T}}\mathbf{S}\mathbf{P}_{\mathrm{v}}\mathbf{b}^{\mathrm{T}}\mathbf{P}_{\mathrm{o}}^{\mathrm{T}} - \mathbf{P}_{\mathrm{o}}\mathbf{b}^{\mathrm{T}}\mathbf{P}_{\mathrm{v}}^{\mathrm{T}}\mathbf{S}\mathbf{P}_{\mathrm{v}}\mathbf{a}\mathbf{P}_{\mathrm{o}}^{\mathrm{T}}) - \\ &(\mathbf{P}_{\mathrm{o}}\mathbf{a}\mathbf{P}_{\mathrm{v}}^{\mathrm{T}}\mathbf{S}\mathbf{P}_{\mathrm{v}}\mathbf{b}^{\mathrm{T}}\mathbf{P}_{\mathrm{o}}^{\mathrm{T}} - \mathbf{P}_{\mathrm{o}}\mathbf{b}^{\mathrm{T}}\mathbf{P}_{\mathrm{v}}^{\mathrm{T}}\mathbf{S}\mathbf{P}_{\mathrm{v}}\mathbf{a}\mathbf{P}_{\mathrm{o}}^{\mathrm{T}})\}\mathbf{S} = \mathbf{0}\end{aligned} \tag{108}$$

## Appendix B. The Exchange-Correlation Contributions to the Property Gradients

According to the equations given in section 2.6, the analytic computation of the geometric derivative properties here discussed requires the geometric derivative of the exchange-correlation energy functional, $\partial E_{\mathrm{xc}}[\rho]/\partial R_\beta$, of the exchange-correlation contribution to the Kohn−Sham matrix, $\partial \mathbf{F}^{\mathrm{xc}}/\partial R_\beta$, as well as the derivative of the exchange-correlation contribution to the (generalized) Kohn−Sham Hessian $\partial \mathbf{G}^{\mathrm{xc}}/\partial R_\beta$. In this appendix, we derive explicit expressions for all of these exchange-correlation derivative contributions.

**B.1. The Geometric Derivative of the Exchange-Correlation Energy Functional, $\partial E_{\mathrm{xc}}[\rho]/\partial R_\beta$.** The exchange-correlation energy $E_{\mathrm{xc}}$ is obtained from integration over space of some functional $f = f[\rho,\nabla\rho]$ of the electron density $\rho = \rho(\mathbf{r})$ and (possibly) of the gradient of the density $\nabla\rho = \nabla\rho(\mathbf{r})$:

$$E_{\mathrm{xc}} = \int f[\rho, \nabla\rho]\, \mathrm{d}\mathbf{r} \tag{109}$$

Differentiation with respect to a nuclear displacement $R$ yields

$$\frac{\partial E_{\mathrm{xc}}[\rho]}{\partial R_\beta} = \int \frac{\delta E_{\mathrm{xc}}[\rho]}{\delta\rho(\mathbf{r})}\frac{\partial\rho(\mathbf{r})}{\partial R_\beta}\, \mathrm{d}\mathbf{r} \tag{110}$$

The derivative of the exchange-correlation energy with respect to the density can be written:[71,72]

$$\frac{\delta E_{\mathrm{xc}}}{\delta\rho} = \frac{\partial f[\rho, \nabla\rho]}{\partial\rho} - \nabla\frac{\partial f[\rho, \nabla\rho]}{\partial\nabla\rho} \equiv \frac{\partial f}{\partial\rho} - \nabla\frac{\partial f}{\partial\nabla\rho} \tag{111}$$

We insert eq 111 into eq 110,

$$\frac{\partial E_{\mathrm{xc}}[\rho]}{\partial R_\beta} = \int \frac{\partial f}{\partial\rho}\frac{\partial\rho(\mathbf{r})}{\partial R_\beta}\, \mathrm{d}\mathbf{r} - \int \frac{\partial\rho(\mathbf{r})}{\partial R_\beta}\left(\nabla\frac{\partial f}{\partial\nabla\rho}\right) \mathrm{d}\mathbf{r} \tag{112}$$

and integrate the second term by parts according to $\int_{-\infty}^{+\infty} u\mathrm{d}v = [uv]_{-\infty}^{+\infty} - \int v\mathrm{d}u$, with $v = \partial f/\partial\nabla\rho$, $\mathrm{d}v = \nabla u$, and $u = \partial\rho(\mathbf{r})/\partial R_\beta$, assuming that the constant term vanishes due to the locality of $f$. We thus obtain

$$\frac{\partial E_{\mathrm{xc}}[\rho]}{\partial R_\beta} = \int \frac{\partial f}{\partial\rho}\frac{\partial\rho(\mathbf{r})}{\partial R_\beta}\, \mathrm{d}\mathbf{r} + \int \frac{\partial f}{\partial\nabla\rho}\frac{\partial\nabla\rho(\mathbf{r})}{\partial R_\beta}\, \mathrm{d}\mathbf{r} \tag{113}$$

It is convenient to make a change of fundamental variable from $\nabla\rho$ to its norm $\xi = \|\nabla\rho\|$ and consequently apply the chain rule $(\partial f(g(x)))/(\partial x) = [(\partial f)/(\partial g)][(\partial g)/(\partial x)]$ in the second term of eq 113 to obtain

$$\frac{\partial E_{\mathrm{xc}}[\rho]}{\partial R_\beta} = \int \frac{\partial f}{\partial\rho}\frac{\partial\rho(\mathbf{r})}{\partial R_\beta}\, \mathrm{d}\mathbf{r} + \int \frac{\partial f}{\partial\xi}\frac{\partial\xi}{\partial\nabla\rho}\frac{\partial\nabla\rho(\mathbf{r})}{\partial R_\beta}\, \mathrm{d}\mathbf{r} \tag{114}$$

From the definition of $\xi$, it follows that

$$\frac{\partial\xi}{\partial\nabla\rho(\mathbf{r})} = \frac{\nabla\rho(\mathbf{r})}{\xi} \tag{115}$$

which may be inserted into eq 114, yielding the geometric derivative (i.e., the gradient) of the exchange-correlation contribution to the ground-state energy

$$\frac{\partial E_{\mathrm{xc}}[\rho]}{\partial R_\beta} = \int \frac{\partial f}{\partial\rho}\frac{\partial\rho(\mathbf{r})}{\partial R_\beta}\, \mathrm{d}\mathbf{r} + \int \frac{\partial f}{\partial\xi}\frac{\nabla\rho(\mathbf{r})}{\xi}\frac{\partial\nabla\rho(\mathbf{r})}{\partial R_\beta}\, \mathrm{d}\mathbf{r} \tag{116}$$

**B.2. The Exchange-Correlation Contribution to the Generalized Kohn−Sham Hessian Matrix, $\mathbf{G}^{\mathrm{xc}}(\mathbf{M})$.** Before deriving the analytic expression of the geometric derivative of the exchange-correlation contribution to the Kohn−Sham matrix, it is convenient to derive the expression of the exchange-correlation contribution to the generalized Kohn−Sham Hessian matrix:

$$G_{\mu\nu}^{\mathrm{xc}}(\mathbf{M}) = \sum_{\rho\sigma} M_{\sigma\rho}\int \frac{\delta^2 E_{\mathrm{xc}}}{\delta\rho(\mathbf{s})\,\delta\rho(\mathbf{r})}\,\Omega_{\mu\nu}(\mathbf{r})\,\Omega_{\rho\sigma}(\mathbf{s})\,\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{s} \tag{117}$$

where $\mathbf{M}$ is a general "perturbed" density matrix, as for instance $\mathbf{D}^b$ in eq 82. Using eq 111, we obtain

$$\begin{aligned}G_{\mu\nu}^{\mathrm{xc}}(\mathbf{M}) = &\sum_{\rho\sigma} M_{\sigma\rho}\int \frac{\delta}{\delta\rho(\mathbf{s})}\frac{\partial f[\mathbf{r}]}{\partial\rho(\mathbf{r})}\Omega_{\mu\nu}(\mathbf{r})\,\Omega_{\rho\sigma}(\mathbf{s})\,\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{s} - \\ &\sum_{\rho\sigma} M_{\sigma\rho}\int \frac{\delta}{\delta\rho(\mathbf{s})}\nabla_r\left(\frac{\partial f[\mathbf{r}]}{\partial\nabla\rho(\mathbf{r})}\right)\Omega_{\mu\nu}(\mathbf{r})\,\Omega_{\rho\sigma}(\mathbf{s})\,\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{s}\end{aligned} \tag{118}$$

where the subscript $r$ in $\nabla_r$ denotes the (electronic) variable with respect to which we differentiate. Using partial integration on the second term and assuming that the constant term vanishes because $f$ is local gives

$$G^{\mathrm{xc}}_{\mu\nu}(\mathbf{M}) = \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\delta}{\delta\rho(\mathbf{s})} \frac{\partial f[\rho(\mathbf{r}), \nabla\rho(\mathbf{r})]}{\partial\rho(\mathbf{r})}\, \Omega_{\mu\nu}(\mathbf{r})\, \Omega_{\rho\sigma}(\mathbf{s})\, \mathrm{d}\mathbf{r}\, \mathrm{d}\mathbf{s} + \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\delta}{\delta\rho(\mathbf{s})} \frac{\partial f[\mathbf{r}]}{\partial\nabla\rho(\mathbf{r})} \nabla_r[\Omega_{\mu\nu}(\mathbf{r})]\Omega_{\rho\sigma}(\mathbf{s})\, \mathrm{d}\mathbf{r}\, \mathrm{d}\mathbf{s} \quad (119)$$

Using the chain rule for functional derivatives

$$\frac{\delta\mathscr{F}}{\delta g(y)} = \int \frac{\delta\mathscr{F}}{\delta f(x)} \frac{\delta f(x)}{\delta g(y)}\, \mathrm{d}x \quad (120)$$

yields

$$G^{\mathrm{xc}}_{\mu\nu}(\mathbf{M}) = P_1 + P_2 + P_3 + P_4 \quad (121)$$

$$P_1 = \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\delta\partial f[\mathbf{r}]}{\delta\rho(\mathbf{t})\partial\rho(\mathbf{r})} \frac{\delta\rho(\mathbf{t})}{\delta\rho(\mathbf{s})}\, \Omega_{\mu\nu}(\mathbf{r})\, \Omega_{\rho\sigma}(\mathbf{s})\, \mathrm{d}\mathbf{r}\, \mathrm{d}\mathbf{s}\, \mathrm{d}\mathbf{t} \quad (122)$$

$$P_2 = \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\delta\partial f[\mathbf{r}]}{\delta\rho(\mathbf{t})\, \partial\nabla\rho(\mathbf{r})} \frac{\delta\rho(\mathbf{t})}{\delta\rho(\mathbf{s})} \nabla_r[\Omega_{\mu\nu}(\mathbf{r})]\, \Omega_{\rho\sigma}(\mathbf{s})\, \mathrm{d}\mathbf{r}\, \mathrm{d}\mathbf{s}\, \mathrm{d}\mathbf{t} \quad (123)$$

$$P_3 = \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\delta\partial f[\mathbf{r}]}{\delta\nabla\rho(\mathbf{t})\, \partial\rho(\mathbf{r})} \frac{\delta\nabla\rho(\mathbf{t})}{\delta\rho(\mathbf{s})} \Omega_{\mu\nu}(\mathbf{r})\, \Omega_{\rho\sigma}(\mathbf{s})\, \mathrm{d}\mathbf{r}\, \mathrm{d}\mathbf{s}\, \mathrm{d}\mathbf{t} \quad (124)$$

$$P_4 = \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\delta\partial f[\mathbf{r}]}{\delta\nabla\rho(\mathbf{t})\, \partial\nabla\rho(\mathbf{r})} \frac{\delta\nabla\rho(\mathbf{t})}{\delta\rho(\mathbf{s})} \nabla_r[\Omega_{\mu\nu}(\mathbf{r})]\, \Omega_{\rho\sigma}(\mathbf{s})\, \mathrm{d}\mathbf{r}\, \mathrm{d}\mathbf{s}\, \mathrm{d}\mathbf{t} \quad (125)$$

Using the relation between functional and standard derivatives for composite functions

$$\frac{\delta\rho(\mathbf{t})}{\delta\rho(\mathbf{s})} = \delta(\mathbf{t} - \mathbf{s}) \quad (126)$$

$$\frac{\delta f[\rho(\mathbf{r})]}{\delta\rho(\mathbf{t})} = \frac{\partial f[\rho(\mathbf{r})]}{\partial\rho(\mathbf{r})} \delta(\mathbf{r} - \mathbf{t}) \quad (127)$$

and integrating first over **s** and then over **t**, the first two terms of eq 121 become

$$P_1 + P_2 = \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\partial^2 f}{\partial\rho^2} \Omega_{\mu\nu}\Omega_{\rho\sigma}\, \mathrm{d}\mathbf{r} + \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\partial^2 f}{\partial\rho\partial\nabla\rho} \nabla(\Omega_{\mu\nu})\, \Omega_{\rho\sigma}\, \mathrm{d}\mathbf{r} \quad (128)$$

Using the relation

$$\frac{\delta\nabla\rho(\mathbf{t})}{\delta\rho(\mathbf{s})} = \nabla_t \frac{\delta\rho(\mathbf{t})}{\delta\rho(\mathbf{s})} = \nabla_t\delta(\mathbf{t} - \mathbf{s}) \quad (129)$$

we obtain for the last two terms

$$P_3 + P_4 = \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\delta\partial f[\mathbf{r}]}{\delta\nabla\rho(\mathbf{t})\, \partial\rho(\mathbf{r})} \times \nabla_t\delta(\mathbf{t} - \mathbf{s})\, \Omega_{\mu\nu}(\mathbf{r})\, \Omega_{\rho\sigma}(\mathbf{s})\, \mathrm{d}\mathbf{r}\, \mathrm{d}\mathbf{s} + \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\delta\partial f[\mathbf{r}]}{\delta\nabla\rho(\mathbf{t})\, \partial\nabla\rho(\mathbf{r})} \times \nabla_t\delta(\mathbf{t} - \mathbf{s})\nabla_r[\Omega_{\mu\nu}(\mathbf{r})]\, \Omega_{\rho\sigma}(\mathbf{s})\, \mathrm{d}\mathbf{r}\, \mathrm{d}\mathbf{s} \quad (130)$$

Integrating by parts—again assuming that the constant term vanishes due to the fact that $f$ is a local function—yields

$$P_3 + P_4 = -\sum_{\rho\sigma} M_{\sigma\rho} \int \nabla_t\Big(\frac{\delta\partial f[\mathbf{r}]}{\delta\nabla\rho(\mathbf{t})\, \partial\rho(\mathbf{r})}\Big) \times \delta(\mathbf{t} - \mathbf{s})\, \Omega_{\mu\nu}(\mathbf{r})\, \Omega_{\rho\sigma}(\mathbf{s})\, \mathrm{d}\mathbf{r}\, \mathrm{d}\mathbf{s}\, \mathrm{d}\mathbf{t} - \sum_{\rho\sigma} M_{\sigma\rho} \times \int \nabla_t\Big(\frac{\delta\partial f[\mathbf{r}]}{\delta\nabla\rho(\mathbf{t})\, \partial\nabla\rho(\mathbf{r})}\Big) \times \delta(\mathbf{t} - \mathbf{s})\nabla_r[\Omega_{\mu\nu}(\mathbf{r})]\, \Omega_{\rho\sigma}(\mathbf{s})\, \mathrm{d}\mathbf{r}\, \mathrm{d}\mathbf{s}\, \mathrm{d}\mathbf{t} \quad (131)$$

Integration over **t** gives

$$P_3 + P_4 = -\sum_{\rho\sigma} M_{\sigma\rho} \int \nabla_s\Big(\frac{\delta\partial f[\mathbf{r}]}{\delta\nabla\rho(\mathbf{s})\, \partial\rho(\mathbf{r})}\Big) \times \Omega_{\mu\nu}(\mathbf{r})\, \Omega_{\rho\sigma}(\mathbf{s})\, \mathrm{d}\mathbf{r}\, \mathrm{d}\mathbf{s} - \sum_{\rho\sigma} M_{\sigma\rho} \int \nabla_s\Big(\frac{\delta\partial f[\mathbf{r}]}{\delta\nabla\rho(\mathbf{s})\, \partial\nabla\rho(\mathbf{r})}\Big) \times \nabla_r[\Omega_{\mu\nu}(\mathbf{r})]\, \Omega_{\rho\sigma}(\mathbf{s})\, \mathrm{d}\mathbf{r}\, \mathrm{d}\mathbf{s} \quad (132)$$

and yet another partial integration and an integration over **s** using eq 127 yield

$$P_3 + P_4 = \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\partial^2 f}{\partial\rho\partial\nabla\rho} \Omega_{\mu\nu}\nabla\Omega_{\rho\sigma}\, \mathrm{d}\mathbf{r} + \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\partial^2 f}{\partial(\nabla\rho)^2} \nabla\Omega_{\mu\nu}\nabla\Omega_{\rho\sigma}\, \mathrm{d}\mathbf{r} \quad (133)$$

The total $G^{\mathrm{xc}}_{\mu\nu}(\mathbf{M})$ may thus be obtained from eqs 128 and 133, giving

$$G^{\mathrm{xc}}_{\mu\nu}(\mathbf{M}) = \sum_{\rho\sigma} M_{\sigma\rho} \int \Big\{ \frac{\partial^2 f}{\partial\rho^2} \Omega_{\mu\nu}\Omega_{\rho\sigma} + \frac{\partial^2 f}{\partial\rho\partial\nabla\rho} (\Omega_{\mu\nu}\nabla\Omega_{\rho\sigma} + \Omega_{\rho\sigma}\nabla\Omega_{\mu\nu}) \Big\}\, \mathrm{d}\mathbf{r} + \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\partial^2 f}{\partial(\nabla\rho)^2} \nabla\Omega_{\mu\nu}\, \nabla\Omega_{\rho\sigma}\, \mathrm{d}\mathbf{r} \quad (134)$$

A transformation of variables from $\nabla\rho$ to $\xi = \|\nabla\rho\|$ yields

$$G^{\mathrm{xc}}_{\mu\nu}(\mathbf{M}) = \sum_{\rho\sigma} M_{\sigma\rho} \int \Big\{ \frac{\partial^2 f}{\partial\rho^2} \Omega_{\mu\nu}\Omega_{\rho\sigma} + \frac{\partial^2 f}{\partial\rho\partial\xi} \Big( \Omega_{\mu\nu} \frac{\nabla\rho\nabla\Omega_{\rho\sigma}}{\xi} + \Omega_{\rho\sigma} \frac{\nabla\Omega_{\mu\nu}\nabla\rho}{\xi} \Big) \Big\}\, \mathrm{d}\mathbf{r} + \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\partial^2 f}{\partial\xi^2} \frac{\nabla\Omega_{\mu\nu}\nabla_\rho}{\xi} \frac{\nabla\Omega_{\rho\sigma}\nabla\rho}{\xi}\, \mathrm{d}\mathbf{r} \quad (135)$$

We can now use the general chain rule for second-order derivatives

$$\frac{\partial^2 f}{\partial x^2} = \frac{\partial^2 f}{\partial g^2}\left(\frac{\partial g}{\partial x}\right)^2 + \frac{\partial f}{\partial g}\frac{\partial^2 g}{\partial x^2} \tag{136}$$

to get

$$\frac{\partial^2 f}{\partial(\nabla\rho)^2} = \frac{\partial^2 f}{\partial\xi^2}\left(\frac{\partial\xi}{\partial\nabla\rho}\right)^2 + \frac{\partial f}{\partial\xi}\frac{\partial^2\xi}{\partial(\nabla\rho)^2} = \frac{\partial^2 f}{\partial\xi^2}\left(\frac{\nabla\rho(\mathbf{r})}{\xi}\right)^2 \tag{137}$$

since

$$\frac{\partial\xi}{\partial\nabla\rho(\mathbf{r})} = \frac{\nabla\rho(\mathbf{r})}{\xi}; \qquad \frac{\partial^2\xi}{\partial[\nabla\rho(\mathbf{r})]^2} = 0 \tag{138}$$

and do a variable transform from $\xi = \|\nabla\rho\|$ to $z = \|\nabla\rho\|^2$ and use again the general chan rule for second-order derivatives to write

$$\frac{\partial^2 f}{\partial\xi^2} = \frac{\partial^2 f}{\partial z^2}\left(\frac{\partial z}{\partial\xi}\right)^2 + \frac{\partial f}{\partial z}\frac{\partial^2 z}{\partial\xi^2} \tag{139}$$

$$\frac{\partial z}{\partial\xi} = 2\xi; \qquad \frac{\partial^2 z}{\partial\xi^2} = 2 \tag{140}$$

and thus obtain

$$\begin{aligned} G^{\mathrm{xc}}_{\mu\nu}(\mathbf{M}) = {} & \sum_{\rho\sigma} M_{\sigma\rho}\int\left\{\frac{\partial^2 f}{\partial\rho^2}\Omega_{\mu\nu}\Omega_{\rho\sigma} + \right. \\ & \left. \frac{\partial^2 f}{\partial\rho\partial z}\,2[\Omega_{\mu\nu}(\nabla\rho\nabla\Omega_{\rho\sigma}) + \Omega_{\rho\sigma}(\nabla\Omega_{\mu\nu}\nabla\rho)]\right\}\mathrm{d}\mathbf{r} + \\ & \sum_{\rho\sigma} M_{\sigma\rho}\int\frac{\partial^2 f}{\partial z^2}4(\nabla\Omega_{\mu\nu}\nabla\rho)\,(\nabla\Omega_{\rho\sigma}\nabla\rho)\,\mathrm{d}\mathbf{r} + \\ & \sum_{\rho\sigma} M_{\sigma\rho}\int\frac{\partial f}{\partial z}2\frac{\nabla\Omega_{\mu\nu}\nabla_\rho}{\xi}\frac{\nabla\Omega_{\rho\sigma}\nabla\rho}{\xi}\,\mathrm{d}\mathbf{r} \end{aligned} \tag{141}$$

Note the differences compared to the expression given in eq 124 of ref 28.

**B.3. The Exchange-Correlation Contribution to the Differentiated Kohn−Sham Matrix, $\partial F^{\mathrm{xc}}/\partial R_\beta$.** The exchange-correlation contribution to the Kohn−Sham matrix is given by (see also eq 8)

$$F^{\mathrm{xc}}_{\mu\nu} = \int\frac{\delta E^{\mathrm{xc}}[\rho]}{\delta\rho(\mathbf{r})}\Omega_{\mu\nu}(\mathbf{r})\,\mathrm{d}\mathbf{r} \tag{142}$$

The derivative of the exchange-correlation contribution to the Kohn−Sham matrix is given by

$$\begin{aligned} \frac{\partial F^{\mathrm{xc}}_{\mu\nu}}{\partial R_\beta} = {} & \int\frac{\delta^2 E^{\mathrm{xc}}[\rho]}{\delta\rho(\mathbf{r})\,\delta\rho(\mathbf{s})}\frac{\partial\rho(\mathbf{s})}{\partial R_\beta}\Omega_{\mu\nu}(\mathbf{r})\,\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{s} + \\ & \int\frac{\delta E^{\mathrm{xc}}[\rho]}{\delta\rho(\mathbf{r})}\frac{\partial\Omega_{\mu\nu}(\mathbf{r})}{\partial R_\beta}\,\mathrm{d}\mathbf{r} \end{aligned} \tag{143}$$

which can be rewritten as

$$\begin{aligned} \frac{\partial F^{\mathrm{xc}}_{\mu\nu}}{\partial R_\beta} = {} & \sum_{\sigma\rho}\int D_{\sigma\rho}\frac{\delta^2 E^{\mathrm{xc}}[\rho]}{\delta\rho(\mathbf{r})\,\delta\rho(\mathbf{s})}\frac{\partial\Omega_{\rho\sigma}(\mathbf{s})}{\partial R_\beta}\Omega_{\mu\nu}(\mathbf{r})\,\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{s} + \\ & \sum_{\sigma\rho}\int\frac{\partial D_{\sigma\rho}}{\partial R_\beta}\frac{\delta^2 E^{\mathrm{xc}}[\rho]}{\delta\rho(\mathbf{r})\delta\rho(\mathbf{s})}\Omega_{\rho\sigma}(\mathbf{s})\,\Omega_{\mu\nu}(\mathbf{r})\,\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{s} + \\ & \int\frac{\delta E^{\mathrm{xc}}[\rho]}{\delta\rho(\mathbf{r})}\frac{\partial\Omega_{\mu\nu}(\mathbf{r})}{\partial R_\beta}\,\mathrm{d}\mathbf{r} \equiv Q_1 + Q_2 + Q_3 \end{aligned} \tag{144}$$

Comparing the second term $Q_2$ with the exchange-correlation contribution to the generalized Hessian in eq 117, we can immediately write

$$Q_2 = G^{\mathrm{xc}}_{\mu\nu}\left(\frac{\partial\mathbf{D}}{\partial R_\beta}\right) \tag{145}$$

The last term $Q_3$ of eq 144 is structurally similar to eq 110 and may be written as

$$Q_3 = \int\frac{\partial f}{\partial\rho}\frac{\partial\Omega_{\mu\nu}}{\partial R_\beta}\,\mathrm{d}\mathbf{r} + \int\frac{\partial f}{\partial\xi}\frac{\nabla\rho}{\xi}\frac{\partial\nabla\Omega_{\mu\nu}}{\partial R_\beta}\,\mathrm{d}\mathbf{r} \tag{146}$$

The first term $Q_1$ is similar to eq 117, except that $\Omega_{\rho\sigma}(\mathbf{s})$ should be replaced by $\partial\Omega_{\rho\sigma}(\mathbf{s})/\partial R_\beta$ and can thus be computed in analogy with terms 134 and 141:

$$\begin{aligned} Q_1 = {} & \sum_{\rho\sigma} D_{\sigma\rho}\int\left\{\frac{\partial^2 f}{\partial\rho^2}\Omega_{\mu\nu}\frac{\partial\Omega_{\rho\sigma}}{\partial R_\beta} + \frac{\partial^2 f}{\partial\rho\partial z}2\times\right. \\ & \left.\left[\Omega_{\mu\nu}\left(\nabla\rho\frac{\partial\nabla\Omega_{\rho\sigma}}{\partial R_\beta}\right) + \frac{\partial\Omega_{\rho\sigma}}{\partial R_\beta}(\nabla\Omega_{\mu\nu}\nabla\rho)\right]\right\}\mathrm{d}\mathbf{r} + \\ & \sum_{\rho\sigma} D_{\sigma\rho}\int\frac{\partial^2 f}{\partial z^2}4(\nabla\Omega_{\mu\nu}\nabla\rho)\left(\frac{\partial\Omega_{\rho\sigma}}{\partial R_\beta}\nabla\rho\right)\mathrm{d}\mathbf{r} + \\ & \sum_{\rho\sigma} D_{\sigma\rho}\int\frac{\partial f}{\partial z^2}2\frac{\nabla\Omega_{\mu\nu}\nabla\rho}{\xi}\frac{\nabla\frac{\partial\Omega_{\rho\sigma}}{\partial R_\beta}\nabla\rho}{\xi}\,\mathrm{d}\mathbf{r} \end{aligned} \tag{147}$$

**B.4. The Additional Exchange-Correlation Contribution to the Quadratic Response $T^{\mathrm{xc}}_{\mu\nu}(\mathbf{N},\mathbf{M})$.** As a last preliminary step to be able to compute the derivative of the exchange-correlation contribution to the generalized Kohn−Sham matrix Hessian, we derive here the explicit expression for the additional exchange-correlation contribution required in computing quadratic response properties:[41]

$$\begin{aligned} T^{\mathrm{xc}}_{\mu\nu}(\mathbf{N},\mathbf{M}) = {} & \sum_{\rho\sigma\eta\varepsilon} M_{\sigma\rho}N_{\varepsilon\eta}\int\Omega_{\eta\varepsilon}(\mathbf{t})\,\Omega_{\rho\sigma}(\mathbf{s})\,\Omega_{\mu\nu}(\mathbf{r})\;\times \\ & \frac{\delta^2 v_{\mathrm{xc}}(\mathbf{r})}{\delta\rho(\mathbf{s})\,\delta\rho(\mathbf{t})}\,\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{s}\,\mathrm{d}\mathbf{t} \end{aligned} \tag{148}$$

where $\mathbf{M}$ and $\mathbf{N}$ are general “perturbed” density matrices, like $\mathbf{D}^b$ in eq 82. Defining

$$\kappa(\mathbf{r}) = \sum_{\sigma\rho} M_{\sigma\rho}\,\Omega_{\rho\sigma}(\mathbf{r}) \tag{149}$$

$$\nabla\kappa(\mathbf{r}) = \sum_{\sigma\rho} M_{\sigma\rho}\nabla\Omega_{\rho\sigma}(\mathbf{r}) \tag{150}$$

$$\tau(\mathbf{r}) = \sum_{\sigma\rho} N_{\sigma\rho}\,\Omega_{\rho\sigma}(\mathbf{r}) \tag{151}$$

$$\nabla\tau(\mathbf{r}) = \sum_{\sigma\rho} N_{\sigma\rho}\nabla\Omega_{\rho\sigma}(\mathbf{r}) \tag{152}$$

we can write, in a shorthand notation,

$$T^{\mathrm{xc}}_{\mu\nu} = \int \tau(\mathbf{t})\,\kappa(\mathbf{s})\,\Omega_{\mu\nu}(\mathbf{r})\,\frac{\delta^2 v_{\mathrm{xc}}(\mathbf{r})}{\delta\rho(\mathbf{s})\,\delta\rho(\mathbf{t})}\,\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{s}\,\mathrm{d}\mathbf{t} \tag{153}$$

The first functional derivative yields

$$\begin{aligned} T^{\mathrm{xc}}_{\mu\nu} = \int \Bigg[&\tau(\mathbf{t})\frac{\delta}{\delta\rho(\mathbf{t})}\frac{\partial^2 f}{\partial\rho^2}\Omega_{\mu\nu}\kappa\ + \\ &\tau(\mathbf{t})\frac{\delta}{\delta\rho(\mathbf{t})}\frac{\partial^2 f}{\partial\rho\partial\nabla\rho}(\Omega_{\mu\nu}\nabla\kappa + \kappa\nabla\Omega_{\mu\nu})\Bigg]\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{t}\ + \\ &\int \tau(\mathbf{t})\frac{\delta}{\delta\rho(\mathbf{t})}\frac{\partial^2 f}{\partial(\nabla\rho)^2}\nabla\Omega_{\mu\nu}\nabla\kappa\,\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{t} \end{aligned} \tag{154}$$

Doing the same for the second functional derivative, we obtain

$$\begin{aligned} T^{\mathrm{xc}}_{\mu\nu} = \int \Bigg(&\frac{\partial^3 f}{\partial\rho^3}\kappa\tau + \frac{\partial^3 f}{\partial\rho^2\partial\nabla\rho}\kappa\nabla\tau + \frac{\partial^3 f}{\partial\rho^2\partial\nabla\rho}\tau\nabla\kappa\ + \\ &\frac{\partial^3 f}{\partial\rho\partial(\nabla\rho)^2}\nabla\kappa\nabla\tau\Bigg)\Omega_{\mu\nu}\,\mathrm{d}\mathbf{r} + \int \frac{\partial^3 f}{\partial\rho^2\partial\nabla\rho}\tau\kappa\nabla\Omega_{\mu\nu}\,\mathrm{d}\mathbf{r}\ + \\ &\int \frac{\partial^3 f}{\partial\rho\partial(\nabla\rho)^2}\nabla\tau\kappa\nabla\Omega_{\mu\nu}\,\mathrm{d}\mathbf{r}\ + \int \frac{\partial^3 f}{\partial\rho\partial(\nabla\rho)^2}\nabla\Omega_{\mu\nu}\nabla\kappa\tau\,\mathrm{d}\mathbf{r}\ + \\ &\int \frac{\partial^3 f}{\partial(\nabla\rho)^3}\nabla\Omega_{\mu\nu}\nabla\kappa\nabla\tau\,\mathrm{d}\mathbf{r} \end{aligned} \tag{155}$$

A variable transformation from $\nabla\rho$ to $\xi = \|\nabla\rho\|$ gives

$$\begin{aligned} T^{\mathrm{xc}}_{\mu\nu} = \int \Bigg(&\frac{\partial^3 f}{\partial\rho^3}\kappa\tau + \frac{\partial^3 f}{\partial\rho^2\partial\xi}\frac{\nabla\rho}{\xi}\kappa\nabla\tau + \frac{\partial^3 f}{\partial\rho^2\partial\xi}\tau\nabla\kappa\frac{\nabla\rho}{\xi}\ + \\ &\frac{\partial^3 f}{\partial\rho\partial\xi^2}\Big(\frac{\nabla\rho}{\xi}\Big)^2\nabla\kappa\nabla\tau\Bigg)\Omega_{\mu\nu}\,\mathrm{d}\mathbf{r} + \int \frac{\partial^3 f}{\partial\rho^2\partial\xi}\frac{\nabla\rho}{\xi}\tau\kappa\nabla\Omega_{\mu\nu}\,\mathrm{d}\mathbf{r}\ + \\ &\int \frac{\partial^3 f}{\partial\rho\partial\xi^2}\Big(\frac{\nabla\rho}{\xi}\Big)^2\nabla\tau\kappa\nabla\Omega_{\mu\nu}\,\mathrm{d}\mathbf{r} + \int \frac{\partial^3 f}{\partial\rho^2\partial\xi}\Big(\frac{\nabla\rho}{\xi}\Big)^2\nabla\Omega_{\mu\nu}\nabla\kappa\tau\,\mathrm{d}\mathbf{r}\ + \\ &\int \frac{\partial^3 f}{\partial\xi^3}\Big(\frac{\nabla\rho}{\xi}\Big)^3\nabla\Omega_{\mu\nu}\nabla\kappa\nabla\tau\,\mathrm{d}\mathbf{r} \end{aligned} \tag{156}$$

From $(\partial f)/(\partial\nabla\rho) = [(\partial f)/(\partial\xi)][(\partial\xi)/(\partial\nabla\rho)] = (\nabla\rho)/(\xi)$ and the general chain rule for second-order derivatives in eq 136, we can write

$$\frac{\partial^3 f}{\partial(\nabla\rho)^3} = \frac{\partial^3 f}{\partial\xi^3}\Big(\frac{\partial\xi}{\partial\nabla\rho}\Big)^3 + 3\frac{\partial^2 f}{\partial\xi^2}\frac{\partial^2\xi}{\partial(\nabla\rho)^2}\frac{\partial\xi}{\partial\nabla\rho} + \frac{\partial f}{\partial\xi}\frac{\partial^3\xi}{\partial(\nabla\rho)^3} = \frac{\partial^3 f}{\partial\xi^3}\Big(\frac{\nabla\rho}{\xi}\Big)^3 \tag{157}$$

since

$$\frac{\partial\xi}{\partial\nabla\rho} = \frac{\nabla\rho(\mathbf{r})}{\xi} \tag{158}$$

$$\frac{\partial^2\xi}{\partial(\nabla\rho)^2} = 0 \tag{159}$$

$$\frac{\partial^3\xi}{\partial(\nabla\rho)^3} = 0 \tag{160}$$

Another variable transformation from $\xi = \|\nabla\rho\|$ to $z = \|\nabla\rho\|^2$ and application of the chain rule for higher-order derivatives gives

$$\frac{\partial f}{\partial\xi} = \frac{\partial f}{\partial z}\frac{\partial z}{\partial\xi} = \frac{\partial f}{\partial z}2\xi \tag{161}$$

$$\frac{\partial^2 f}{\partial\xi^2} = \frac{\partial^2 f}{\partial z^2}\Big(\frac{\partial z}{\partial\xi}\Big)^2 + \frac{\partial f}{\partial z}\frac{\partial^2 z}{\partial\xi^2} = \frac{\partial^2 f}{\partial z^2}4\xi^2 + \frac{\partial f}{\partial z}2 \tag{162}$$

$$\frac{\partial^3 f}{\partial\xi^3} = \frac{\partial^3 f}{\partial z^3}\Big(\frac{\partial z}{\partial\xi}\Big)^3 + 3\frac{\partial^2 f}{\partial z^2}\frac{\partial^2 z}{\partial\xi^2}\frac{\partial z}{\partial x} + \frac{\partial f}{\partial z}\frac{\partial^3 z}{\partial\xi^3} = \frac{\partial^3 f}{\partial z^3}8\xi^3 + 3\frac{\partial^2 f}{\partial z^2}4\xi \tag{163}$$

$$\frac{\partial z}{\partial\xi} = 2\xi \tag{164}$$

$$\frac{\partial^2 z}{\partial\xi^2} = 2 \tag{165}$$

and we obtain

$$\begin{aligned} T^{\mathrm{xc}}_{\mu\nu} = \int \Bigg(&\frac{\partial^3 f}{\partial\rho^3}\kappa\tau + \frac{\partial^3 f}{\partial\rho^2\partial z}2\nabla\rho\kappa\nabla\tau + \frac{\partial^3 f}{\partial\rho^2\partial z}\tau\nabla\kappa 2\nabla\rho\ + \\ &\frac{\partial^3 f}{\partial\rho\partial z^2}4(\nabla\rho)^2\nabla\kappa\nabla\tau + \frac{\partial^2 f}{\partial\rho\partial z}2\Big(\frac{\nabla\rho}{\xi}\Big)^2\nabla\kappa\nabla\tau\Bigg)\Omega_{\mu\nu}\,\mathrm{d}\mathbf{r}\ + \\ &\int \frac{\partial^3 f}{\partial\rho^2\partial z}2\nabla\rho\tau\kappa\nabla\Omega_{\mu\nu}\,\mathrm{d}\mathbf{r}\ + \\ &\int \frac{\partial^3 f}{\partial\rho\partial z^2}4(\nabla\rho)^2\nabla\tau\kappa\nabla\Omega_{\mu\nu}\,\mathrm{d}\mathbf{r}\ + \\ &\int \frac{\partial^2 f}{\partial\rho\partial z}\,2\Big(\frac{\nabla\rho}{\xi}\Big)^2\nabla\tau\kappa\nabla\Omega_{\mu\nu}\,\mathrm{d}\mathbf{r}\ + \\ &\int \frac{\partial^3 f}{\partial\rho\partial z^2}4(\nabla\rho)^2\nabla\Omega_{\mu\nu}\nabla\kappa\tau\,\mathrm{d}\mathbf{r}\ + \\ &\int \frac{\partial^2 f}{\partial\rho\partial z}2\Big(\frac{\nabla\rho}{\xi}\Big)^2\nabla\Omega_{\mu\nu}\nabla\kappa\tau\,\mathrm{d}\mathbf{r}\ + \\ &\int \frac{\partial^3 f}{\partial z^3}8(\nabla\rho)^3\nabla\Omega_{\mu\nu}\nabla\kappa\nabla\tau\,\mathrm{d}\mathbf{r}\ + \\ &\int \frac{\partial^2 f}{\partial z^2}3\cdot 4\xi\Big(\frac{\nabla\rho}{\xi}\Big)^3\nabla\Omega_{\mu\nu}\nabla\kappa\nabla\tau\,\mathrm{d}\mathbf{r} \end{aligned} \tag{166}$$

### B.5. The Derivative of the Kohn−Sham Contribution to the Generalized Hessian $\partial G^{\mathrm{xc}}_{\mu\nu}(\mathbf{M})/\partial R_\beta$.

$$\begin{aligned}
\frac{\partial G^{\mathrm{xc}}_{\mu\nu}(\mathbf{M})}{\partial R_\beta} = {} & \sum_{\rho\sigma} \frac{\partial M_{\sigma\rho}}{\partial R_\beta} \int \frac{\delta^2 E^{\mathrm{xc}}}{\delta\rho(\mathbf{s})\,\delta\rho(\mathbf{r})} \times \\
& \Omega_{\mu\nu}(\mathbf{r})\,\Omega_{\rho\sigma}(\mathbf{s})\,\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{s} \\
& + \sum_{\rho\sigma\eta\varepsilon} M_{\sigma\rho} \int \frac{\delta^3 E^{\mathrm{xc}}}{\delta\rho(\mathbf{t})\,\delta\rho(\mathbf{s})\,\delta\rho(\mathbf{r})} \frac{\partial D_{\varepsilon\eta}}{\partial R_\beta} \times \\
& \Omega_{\eta\varepsilon}(\mathbf{t})\,\Omega_{\mu\nu}(\mathbf{r})\,\Omega_{\rho\sigma}(\mathbf{s})\,\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{s}\,\mathrm{d}\mathbf{t} \\
& + \sum_{\rho\sigma\varepsilon\eta} D_{\varepsilon\eta} M_{\sigma\rho} \int \frac{\delta^3 E^{\mathrm{xc}}}{\delta\rho(\mathbf{t})\,\delta\rho(\mathbf{s})\,\delta\rho(\mathbf{r})} \frac{\partial \Omega_{\eta\varepsilon}(\mathbf{t})}{\partial R_\beta} \times \\
& \Omega_{\mu\nu}(\mathbf{r})\,\Omega_{\rho\sigma}(\mathbf{s})\,\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{s}\,\mathrm{d}\mathbf{t} \\
& + \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\delta^2 E^{\mathrm{xc}}}{\delta\rho(\mathbf{s})\,\delta\rho(\mathbf{r})} \frac{\partial \Omega_{\mu\nu}(\mathbf{r})}{\partial R_\beta} \times \\
& \Omega_{\rho\sigma}(\mathbf{s})\,\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{s} \\
& + \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\delta^2 E^{\mathrm{xc}}}{\delta\rho(\mathbf{s})\,\delta\rho(\mathbf{r})} \Omega_{\mu\nu}(\mathbf{r}) \frac{\partial \Omega_{\rho\sigma}(\mathbf{s})}{\partial R_\beta}\,\mathrm{d}\mathbf{r}\,\mathrm{d}\mathbf{s} \\
= {} & Z_1 + Z_2 + Z_3 + Z_4 + Z_5
\end{aligned} \tag{167}$$

From consideration of the structure of the five terms above, it may be seen that

$$Z_1 = G^{\mathrm{xc}}_{\mu\nu}\left(\frac{\partial \mathbf{M}}{\partial R_\beta}\right) \tag{168}$$

$$Z_2 = T^{\mathrm{xc}}_{\mu\nu}\left(\frac{\partial \mathbf{D}}{\partial R_\beta}, \mathbf{M}\right) \tag{169}$$

The term $Z_3$ is similar to eq 59 (i.e., eq 148) except that $\Omega_{\eta\epsilon}(\mathbf{t})$ should be replaced with $[\partial\Omega_{\eta\epsilon}(\mathbf{t})]/(\partial R_\beta)$. The terms $Z_4$ and $Z_5$ are similar to eq 117 except that $\Omega_{\mu\nu}(\mathbf{r})$ and $\Omega_{\rho\sigma}(\mathbf{s})$ should be replaced with $[\partial\Omega_{\mu\nu}(\mathbf{r})]/(\partial R_\beta)$ and $[\partial\Omega_{\rho\sigma}(\mathbf{s})]/(\partial R_\beta)$, respectively.

## References

(1) Helgaker, T. *Int. J. Quantum Chem.* **1982**, *21*, 939–940.

(2) Bartlett, R. J. In *Geometrical Derivatives of Energy Surfaces and Molecular Properties*; Jørgensen, P., Simons, J., Eds.; Reidel: Dordrecht, The Netherlands, 1986.

(3) Helgaker, T.; Ruud, K.; Bak, K. L.; Jørgensen, P.; Olsen, J. *Faraday Discuss.* **1994**, *99*, 165.

(4) Pulay, P. In *Modern Electronic Structure Theory, Part II*; Yarkony, D. R., Ed.; World Scientific: Singapore, 1995.

(5) Quinet, O.; Champagne, B. *J. Chem. Phys.* **2001**, *115* (14), 6293–6299.

(6) Quinet, O.; Champagne, B.; Kirtman, B. *J. Comput. Chem.* **2001**, *22*, 1920–1932.

(7) Quinet, O.; Champagne, B.; Kirtman, B. *J. Chem. Phys.* **2003**, *117* (6), 2481–2488.

(8) Liégeois, V.; Ruud, K.; Champagne, B. *J. Chem. Phys.* **2007**, *127* (1−6), 204105.

(9) Rappoport, D.; Furche, F. *J. Chem. Phys.* **2007**, *126*, 201104.

(10) Barron, L. D. *Molecular Light Scattering and Optical Activity*, 2nd ed. revised and enlarged; Cambridge University Press: Cambridge, U. K., 2004.

(11) Thorvaldsen, A. J.; Ferrighi, L.; Ruud, K.; Ågren, H.; Coriani, S.; Jørgensen, P. *Phys. Chem. Chem. Phys.* **2009**, *11*, 2293–2304.

(12) Ruud, K.; Thorvaldsen, A. *Chirality* **2009**, *21*, S54–S67.

(13) Berger, R.; Fisher, C.; Klessinger, M. *J. Phys. Chem. A* **1998**, *112*, 7157–7167.

(14) Lin, N.; Zhao, X.; Rizzo, A.; Luo, Y. *J. Chem. Phys.* **2007**, *127*, 244509.

(15) Santoro, F.; Lami, A.; Improta, R.; Bloino, J.; Barone, V. *J. Chem. Phys.* **2008**, *128*, 224311.

(16) Helgaker, T.; Jørgensen, P. Calculation of Geometrical Derivatives in Molecular Electronic Structure Theory. In *Methods in Computational Molecular Physics*; Wilson, S., Diercksen, G. H. F., Eds.; Plenum Press: New York, 1992.

(17) Helgaker, T.; Jørgensen, P. Analytical Calculation of Geometrical Derivatives in Molecular Electronic Structure Theory. In *Adv. Quantum Chem.*; Academic Press: New York, 1988; Vol 19.

(18) Jørgensen, P.; Helgaker, T. *J. Chem. Phys.* **1988**, *89*, 1560–1570.

(19) Helgaker, T.; Jørgensen, P. *Theor. Chim. Acta* **1989**, *75*, 111–127.

(20) Helgaker, T.; Jørgensen, P.; Handy, N. *Theor. Chim. Acta* **1989**, *76*, 227–245.

(21) Sasagane, K.; Aiga, F.; Itoh, R. *J. Chem. Phys.* **1993**, *99*, 3738.

(22) Christiansen, O.; Hättig, C.; Jørgensen, P. *Int. J. Quantum Chem.* **1998**, *68*, 1.

(23) Furche, F.; Ahlrichs, R. *J. Chem. Phys.* **2002**, *117*, 7433.

(24) Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry. Introduction to Advanced Electronic Structure Theory*; Dover Publications: Mineola, NY, 1996.

(25) Helgaker, T.; Jørgensen, P.; Olsen, J. *Molecular Electronic-Structure Theory*; Wiley: Chichester, U. K., 2000.

(26) Koch, H.; Jensen, H. J. A.; Jørgensen, P.; Helgaker, T.; Scuseria, G. E.; Schaefer, H. F., III. *Phys. Chem. Chem. Phys.* **2009**, *11*, 2293–2304.

(27) Hald, K.; Halkier, A.; Jørgensen, P.; Coriani, S.; Hättig, C.; Helgaker, T. *J. Chem. Phys.* **2003**, *118*, 2985.

(28) Kjærgaard, T.; Jørgensen, P.; Thorvaldsen, A. J.; Salek, P.; Coriani, S. *J. Chem. Theory Comput.* **2009**, *5*, 1997–2020.

(29) Hättig, C.; Christiansen, O.; Jørgensen, P. *J. Chem. Phys.* **1998**, *108*, 8331–8354.

(30) Hättig, C.; Jørgensen, P. *J. Chem. Phys.* **1998**, *109*, 9219.

(31) Sałek, P.; Vahtras, O.; Helgaker, T.; Ågren, H. *J. Chem. Phys.* **2002**, *117*, 9630.

(32) Thorvaldsen, A.; Ruud, K.; Kristensen, K.; Jørgensen, P.; Coriani, S. *J. Chem. Phys.* **2008**, *129*, 214108.

(33) Larsen, H.; Jørgensen, P.; Olsen, J.; Helgaker, T. *J. Chem. Phys.* **2000**, *113*, 8908.

(34) Coriani, S.; Høst, S.; Jansík, B.; Thøgersen, L.; Olsen, J.; Jørgensen, P.; Reine, S.; Pawlowski, F.; Helgaker, T.; Sałek, P. *J. Chem. Phys.* **2007**, *126*, 154108.

(35) Helgaker, T.; Larsen, H.; Olsen, J.; Jørgensen, P. *Chem. Phys. Lett.* **2000**, *327*, 397.

(36) Coriani, S.; Hättig, C.; Jørgensen, P.; Helgaker, T. *J. Chem. Phys.* **2000**, *113*, 3561.

(37) Kjærgaard, T.; Jansík, B.; Jørgensen, P.; Coriani, S.; Michl, J. *J. Phys. Chem. A* **2007**, *111*, 11278–11286.

(38) Thorvaldsen, A.; Ruud, K.; Jaszunski, M. *J. Phys. Chem. A* **2008**, *112*, 11942.

(39) Bernhardsson, A.; Forsberg, N.; Malmqvist, P.-Å.; Roos, B. *J. Chem. Phys.* **2000**, *112*, 2798–2809.

(40) Larsen, H.; Helgaker, T.; Jørgensen, P.; Olsen, J. *J. Chem. Phys.* **2001**, *115*, 10344.

(41) Kjærgaard, T.; Jørgensen, P.; Olsen, J.; Coriani, S.; Helgaker, T. *J. Chem. Phys.* **2008**, *129*, 054106.

(42) Olsen, J.; Jørgensen, P. Time-Dependent Response Theory with Applications to Self-Consistent Field and Multiconfigurational Self-Consistent Field Wave Functions. In *Modern Electronic Structure Theory, Part II*; Yarkony, D. R., Ed.; World Scientific: Singapore, 1995.

(43) Hettema, H.; Jensen, H. J. A.; Jørgensen, P.; Olsen, J. *J. Chem. Phys.* **1992**, *97*, 1174.

(44) Helgaker, T. U.; Jensen, H.; Jørgensen, P. *J. Chem. Phys.* **1986**, *182*, 6280.

(45) DALTON, an ab initio electronic structure program, Release 2.0, 2005. See http://www.kjemi.uio.no/software/dalton/dalton.html.

(46) Sałek, P.; Høst, S.; Thøgersen, L.; Jørgensen, P.; Manninen, P.; Olsen, J.; Jansík, B.; Reine, S.; Pawlowski, F.; Tellgren, E.; Helgaker, T.; Coriani, S. *J. Chem. Phys.* **2007**, *126*, 114110.

(47) Cohen, E.; Cvitas, T.; Frey, J.; Holmstrom, B.; Kuchitsu, K.; Marquardt, R.; Mills, I.; Pavese, F.; Quack, M.; Stohner, J.; Strauss, H.; Takami, M.; Thor, A. *Quantities, Units, and Symbols in Physical Chemistry*, 3rd ed., also known as the IUPAC Green Book; RSC Publishing: London, U. K., 2007.

(48) Crawford, B., Jr. *J. Chem. Phys.* **1958**, *29*, 1042–1045.

(49) Orlandi, G.; Siebrand, W. *Chem. Phys. Lett.* **1972**, *15*, 465–468.

(50) Huh, J. S.; Stuber, J. L.; Berger, R. Vibronic Transitions in Large Molecular Systems: The Thermal Time-Correlation Function and Rigorous Prescreening of Herzberg-Teller Terms. To be published.

(51) Jankowiak, H.-C.; Stuber, J. L.; Berger, R. *J. Chem. Phys.* **2007**, *127*, 234101.

(52) Huh, J. S.; Stuber, J. L.; Berger, R. Vibronic Transitions in Large Molecular Systems: Prescreening Conditions for Franck−Condon Factors at Finite Temperature and the Thermal Time-Correlation Function. To be published.

(53) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(54) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623–11627.

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

(56) Schäfer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829–5835.

(57) Lindh, R.; Malmqvist, P.-Å.; Gagliardi, L. *Theor. Chem. Acc.* **2001**, *106*, 178–187.

(58) Lebedev, V. I. *Zh. Vychisl. Mat. Mat. Fiz.* **1975**, *15*, 48.

(59) Lebedev, V. I. *Zh. Vychisl. Mat. Mat. Fiz.* **1976**, *16*, 293.

(60) Lebedev, V. I. *Sibirsk. Mat. Zh.* **1977**, *18*, 132.

(61) Frigo, M.; Johnson, S. G. *Proc. IEEE* **2005**, *93* (2), 216–231, Special issue on “Program Generation, Optimization, and Platform Adaptation.” .

(62) Wilson, E. B. *Phys. Rev.* **1934**, *45*, 707–711.

(63) Callomon, J.; Dunn, T.; Mills, I. *Trans. R. Soc., London* **1966**, *259*, 499–532.

(64) Christiansen, O.; Stanton, J.; Gauss, J. *J. Chem. Phys.* **1998**, *108*, 3987.

(65) Stephenson, T.; Radloff, P.; Rice, S. *J. Chem. Phys.* **1984**, *81*, 1060–1072.

(66) Fischer, G.; Jakobson, S. *Mol. Phys.* **1979**, *38*, 299–308.

(67) Hiraya, A.; Shobatake, K. *J. Chem. Phys.* **1991**, *94*, 7700.

(68) Pantos, E.; Philis, J.; Bolovinos, A. *J. Mol. Spectrosc.* **1978**, *72*, 36.

(69) Page, R.; Shen, Y.; Lee, Y. *J. Chem. Phys.* **1988**, *88*, 5362–5376.

(70) Radle, W.; Beck, C. *J. Chem. Phys.* **1940**, *7*, 507–513.

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

(72) Sałek, P.; Hesselmann, A. *J. Comput. Chem.* **2007**, *28*, 2569–2575.

CT900506C

JCTC Journal of Chemical Theory and Computation

# Understanding Selectivity of Hard and Soft Metal Cations within Biological Systems Using the Subvalence Concept. 1. Application to Blood Coagulation: Direct Cation−Protein Electronic Effects versus Indirect Interactions through Water Networks

B. de Courcy,[†,‡] L. G. Pedersen,[§] O. Parisel,[†,‡] N. Gresh,[||] B. Silvi,[†,‡] J. Pilmé,[†,‡,⊥] and J.-P. Piquemal*[,†,‡]

*UPMC Univ Paris 06, UMR 7616, Laboratoire de Chimie Théorique, case courrier 137, 4 place Jussieu, F-75005, Paris, France, CNRS, UMR 7616, Laboratoire de Chimie Théorique, case courrier 137, 4 place Jussieu, F-75005, Paris, France, Laboratory of Structural Biology, National Institute of Environmental Health, Sciences, Research Triangle Park, North Carolina 27709, Laboratoire de Pharmacochimie Moléculaire et Cellulaire, U648 INSERM, UFR Biomédicale, Université Paris Descartes, 45, rue des Saints-Pères, 75006, Paris, France, Université de Lyon, Université Lyon 1, Faculté de Pharmacie, F-69373 Lyon, Cedex 08, France*

Received October 13, 2009

**Abstract:** Following a previous study by de Courcy et al. (*Interdiscip. Sci. Comput. Life Sci.* **2009**, *1*, 55−60), we demonstrate in this contribution, using quantum chemistry, that metal cations exhibit a specific topological signature in the electron localization of their density interacting with ligands according to their "soft" or "hard" character. Introducing the concept of metal cation subvalence, we show that a metal cation can split its outer-shell density (the so-called subvalent domains or basins) according to it capability to form a partly covalent bond involving charge transfer. Such behavior is investigated by means of several quantum chemical interpretative methods encompasing the topological analysis of the Electron Localization Function (ELF) and Bader's Quantum Theory of Atoms in Molecules (QTAIM) and two energy decomposition analyses (EDA), namely, the Reduced Variational Space (RVS) and Constrained Space Orbital Variations (CSOV) approaches. Further rationalization is performed by computing ELF and QTAIM local properties such as electrostatic distributed moments and local chemical descriptors such as condensed Fukui functions and dual descriptors. These reactivity indexes are computed within the ELF topological analysis in addition to QTAIM offering access to a nonatomic reactivity local index, for example, on lone pairs. We apply this "subvalence" concept to study the cation selectivity in enzymes involved in blood coagulation (GLA domains of three coagulation factors). We show that the calcium ions are clearly able to form partially covalent charge transfer networks between the subdomain of the metal ion and the carboxylate oxygen lone pairs, whereas magnesium does not have such ability. Our analysis also explains the different role of two groups (high affinity and low affinity cation binding sites) present in GLA domains. If the presence of Ca(II) is mandatory in the central "high affinity" region to conserve a proper folding and a charge transfer network, external sites are better stabilized by Mg(II), rather than Ca(II), in agreement with the experiment. The central role of discrete water molecules is also discussed in order to understand the stabilities of the observed X-ray structures of the GLA domain. Indeed, the presence of explicit water molecules generating indirect cation−protein interactions through water networks is shown to be able to reverse the observed electronic selectivity occurring when cations directly interact with the Gla domain without the need of water.

10.1021/ct100089s © 2010 American Chemical Society
Published on Web 03/12/2010

## Introduction

Metal cations play a critical role in many biological systems. In most cases, they are specific in their ability to bind to proteins and thereby confer appropriate biological function or activity. For example, the presence of calcium cations is important in blood clotting, signal transduction, and cell division. Specifically, in the case of blood clotting, it has been experimentally observed that the presence of calcium is required for clot formation.[1] Indeed, calcium cations directly participate in the binding and folding of the $\gamma$-carboxyglutamic acid (Gla)-rich domain that is common to the vitamin-K-dependent serine proteases present in the blood coagulation cascade.[1] Blood plasma does not coagulate in the sole presence of magnesium ions,[2,3] an effect attributed to the concomitant lack of binding of the Gla residues to negatively charged phospholipids when only magnesium ions are present. More precisely, recent X-ray crystal structures with a mixture of both Ca(II) and Mg(II) show that the N-terminus $\omega$-loop segment that is thought to be the key determinant for the binding of the GLA domain to membranes has a disordered structure when only magnesium ions are present.[4] In the presence of calcium ions, GLA domains have been crystallized and a strong Gla-calcium network has been observed with varying degrees of calcium ion coordination. However, despite the mandatory presence of calcium ions needed to structure the GLA domain and allowing it to point the three hydrophobic (or anionic) residues forming "the keel" in the direction of negatively charged phospholipids found in cellular membranes, it has also been experimentally demonstrated that the presence of magnesium ions in addition to calcium enhances the affinity of the enzyme for both the membrane and cofactors.[5]

The ability of the calcium ion to coordinate water molecules with flexible coordination is thought to be important for this function. At the theoretical level, recent first-principle Car−Parrinello[6] and force-field simulation[7] studies have reported an in-depth description of the hydration shells and of the preferred coordination numbers for the calcium and magnesium cations. Although we know that the residence time for water on a magnesium ion is substantially longer than that for a calcium ion,[8] we as yet do not know the physical origins of the differences between the binding of calcium and magnesium ions in biological systems, as *no clear electron structure−biological activity relation has been uncovered in realistic model systems*. In this context, we proposed to apply modern quantum chemistry to study the nature of the binding of calcium and magnesium ions in model systems of factors VII, IX, and X, the structures of which are based on protein structures extracted from the Protein Data Bank.

## Outline

In this work, we will first focus on the fundamental interactions occurring in such systems between the metal cations and their environment. Indeed, thanks to X-ray studies, we know that interactions occur between the calcium or magnesium cations and carboxylate moieties. We will then present an extensive ab initio study of the binding of several hard and soft metal cations to carboxylates in different position. Then, applying quantum topological approaches such as Atoms in Molecules (QTAIM)[9] theory and the topological analysis of the Electron Localization Function (ELF),[10] we shed light on the origin of the different behavior of the electronic structure of metal cation and ligands so as to unravel the specific cation topological signature. The observed topological descriptions will be complemented by intermolecular interaction energy decomposition using the Reduced Variational Space approach (RVS)[11a] and the Constrained-Space Orbital variation (CSOV),[11b] which provide insights about the nature, covalent or electrostatic, of the bonding between the cations and carboxylate. To connect this work to conceptual Density Functional Theory (DFT), we will also provide a detailed analysis by means of local chemical descriptors such as the condensed Fukui functions. In a second part, we will present a study of models of the GLA domain from factors IX, VII, and X of the blood coagulation process using ELF computations complemented by multimolecular RVS energy decomposition analyses.

## Method

**A. Topological Analysis of ELF.** The topological analysis relies on a partition of the molecular space achieved in the framework of the theory of gradient dynamics applied to a scalar potential function, say $V(\mathbf{r})$, called "potential function", which contains the physical or chemical information. This partitioning gives rise to a set of nonoverlapping molecular volumes called basins localized around the maxima of the ELF (the attractors of the vector field). The boundaries between these basins, the separatrices, are zero-flux surfaces satisfying the following condition that every point $\mathbf{r}$ is a unit vector normal to the surface. In the QTAIM theory of Bader,[9] the scalar function is the electron density distribution whose basins have their attractors located on the nuclei and which are therefore associated with the atoms that constitute the molecule. In order to recover a chemist's representation of a molecule consistent with Lewis's valence picture, one must use another "local" function that is able to describe the electron pair regions. For almost two decades, the topological analysis of the ELF has been extensively developed and used to analyze chemical bonding and to investigate chemical reactivity (for reviews, see refs 12 and 13). The ELF can be interpreted as a signature of the electron pair distribution. The relationship of the kernel of ELF to pair functions has been established,[13] but in contrast to these latter, the ELF values are confined in the [0,1] range by a Lorentzian transformation which facilitates the interpretation. The basins

* Corresponding author e-mail: jpp@lct.jussieu.fr.

† UPMC Univ Paris 06.

‡ CNRS, UMR 7616.

§ National Institute of Environmental Health Sciences.

|| Université Paris Descartes.

⊥ Université Lyon 1.

of the ELF are either core basins, labeled C(A) corresponding to the inner shells of atom A and encompassing its nucleus (if Z > 2), or valence basins denoted by V(A,B,C...), where A, B, and C are the element symbols. A valence basin can belong to a single atomic valence shell—in this case, V(A) corresponds to a lone pair—or be shared by several atoms and associated to a bond V(A,B).

The ELF basins closely match the electronic domains of the VSEPR[12] model, and it has been shown that the interbasin repulsion provides a map onto the Gillespie–Nyholm rules which describe molecular geometry.[14,15] It has been recently shown that non-VSEPR structures which occur around neutral atoms belonging to the fourth and higher periods can be explained by considering the structure of the external core shell basins[16] and are hereafter referred to as subvalence basins. Details about the ELF analysis can also be found in a recent review paper dealing with the application of ELF to systems of biological interest.[17]

**B. Integration of Local Properties within the ELF or QTAIM Partition.** *1. Local Electrostatic Moments.* From a quantitative point of view, a population analysis can be carried out by integrating the electron density distribution over the basin volumes. Recently, the distributed moments analysis based on the QTAIM partition[9] and on the ELF basins (DEMEP)[18a] has been introduced, which enables an extended discussion on the nature of bonding in molecules. In this paper, we use more specifically the first moment (denoted as $M_1$), which represents the local dipolar polarization of the density.

That way, the Distributed Electrostatic Moments based on the ELF Partition (DEMEP) allows the calculation of local moments located at nonatomic centers such as lone pairs, $\sigma$ bonds, and $\pi$ systems. Local dipole contributions have been shown to be useful in rationalizing inductive polarization effects and typical hydrogen bond interactions. Moreover, bond quadrupole polarization moments being related to a $\pi$ character enable a discussion of bond multiplicities and sorting of the families of molecules according to their bond order.

To summarize, the $M_0(\Omega)$ monopole term corresponds to the negative of the population (denoted $N$):

$$M_0(\Omega) = -\int_\Omega \rho(\mathbf{r})\,d\tau = -N(\Omega) \tag{1}$$

The first moments or dipolar polarization components of the charge distribution are defined by three-dimensional integrals for a given basin $\Omega$ according to

$$\begin{aligned} M_{1,x}(\Omega) &= -\int_\Omega (x - X_c)\rho(\mathbf{r})\,d\tau \\ M_{1,y}(\Omega) &= -\int_\Omega (y - Y_c)\rho(\mathbf{r})\,d\tau \\ M_{1,z}(\Omega) &= -\int_\Omega (z - Z_c)\rho(\mathbf{r})\,d\tau \end{aligned} \tag{2}$$

where $X_c$, $Y_c$, and $Z_c$ are the Cartesian coordinates of the basin centers.

The five second-moment spherical tensor components can also be calculated and are defined as the quadrupolar polarization terms. They can be seen as the ELF basin equivalents to the atomic quadrupole moments introduced by Popelier[9c] in the case of an QTAIM analysis:

$$\begin{aligned} M_{2,zz}(\Omega) &= -\frac{1}{2}\int_\Omega [3(z - Z_c)^2 - \mathbf{r}^2]\rho(\mathbf{r})\,d\tau \\ M_{2,x^2-y^2}(\Omega) &= -\frac{\sqrt{3}}{2}\int_\Omega [(x - X_c)^2 - (y - Y_c^2)]\rho(\mathbf{r})\,d\tau \\ M_{2,xy}(\Omega) &= -\sqrt{3}\int_\Omega (x - X_c)(y - Y_c)\rho(\mathbf{r})\,d\tau \\ M_{2,xz}(\Omega) &= -\sqrt{3}\int_\Omega (x - X_c)(z - Z_c)\rho(\mathbf{r})\,d\tau \\ M_{2,yz}(\Omega) &= -\sqrt{3}\int_\Omega (y - Y_c)(z - Z_c)\rho(\mathbf{r})\,d\tau \end{aligned} \tag{3}$$

The first- or second-moment basin magnitude is then defined as the square root of the sum of squared components:

$$|\mathbf{M}(\Omega)| = \sqrt{\sum_i M_i(\Omega)^2} \tag{4}$$

Thanks to the invariance of the magnitude of any multipole rank (|M1| or |M2|) with respect to the axis for a given bond or lone pair, the approach allows us to compare the dipolar or quadrupolar polarization of a given basin in different chemical environments.

*2. Fukui Functions as Local Chemical Descriptors.* Beyond the computations of local distributed electrostatic moments, it is also possible to access the topological partition of local chemical descriptors. Among the numerous chemical indicators, the Fukui functions,[18b,c] based on the relative properties of the Highest Occupied Molecular Orbital (HOMO) and the Lowest Unoccupied Molecular Orbital (LUMO), are interesting as they are particularly useful for the interpretation of chemical reactivity, particularly toward nucleophiles or electrophiles.[18d] Indeed, following Parr and Yang, conceptual DFT provides such functions defined in terms of the variation of the chemical potential with respect to changes in the external potential $v(\mathbf{r})$ or equivalently as the derivative of the electron density with respect to changes in the number of electrons $N$.

$$f(\mathbf{r}) = \left[\frac{\delta\mu}{\delta v(\mathbf{r})}\right]_N = \left[\frac{\partial\rho(\mathbf{r})}{\partial N}\right]_{v(\mathbf{r})} \tag{5}$$

Three Fukui functions are usually evaluated: $f^+(r)$, $f^-(r)$, and $f^0(r)$

$$\begin{aligned} f^+(\mathbf{r}) &= \left(\frac{\partial\rho(\mathbf{r})}{\partial N}\right)^+_{v(\mathbf{r})} \approx \left(\frac{\delta E_{\text{LUMO}}}{\delta v(\mathbf{r})}\right)_N \approx \rho_{\text{LUMO}}(\mathbf{r}) \\ f^-(\mathbf{r}) &= \left(\frac{\partial\rho(\mathbf{r})}{\partial N}\right)^-_{v(\mathbf{r})} \approx \left(\frac{\delta E_{\text{HOMO}}}{\delta v(\mathbf{r})}\right)_N \approx \rho_{\text{HOMO}}(\mathbf{r}) \\ f^0(\mathbf{r}) &= \frac{1}{2}[f^+(\mathbf{r}) + f^-(\mathbf{r})] \end{aligned} \tag{6}$$

They are sometime also associated with the computation of another value called the dual descriptor (denoted $\Delta f(r)$[18e,f]) and calculated upon the $f^+(r)$ and the $f^-(r)$ functions:

$$\Delta f(\mathbf{r}) = (f^+(\mathbf{r}) - f^-(\mathbf{r}))_N \tag{7}$$

The $f^+(r)$ function usually characterizes the reactivity of a given species toward nucleophilic attack (in that case, $\Delta f(r) > 0$), whereas the $f^-(r)$ function usually characterizes the reactivity of a given species toward electrophilic attack (in that case, $\Delta f(r) < 0$). When $\Delta f(r) = 0$, the site's reactivity should be equilibrated. As an analytic expression of such function is not available, it remains possible to compute them numerically using finite differences (see for example ref 18g and references therein).

In that case, $f^-(r)$, $f^+(r)$, and $f^0(r)$ can be computed as follows:

$$\begin{aligned} f^-(\mathbf{r}) &= [q_x(N) - q_x(N-1)] \\ f^+(\mathbf{r}) &= [q_x(N+1) - q_x(N)] \\ f^0(\mathbf{r}) &= \frac{1}{2}[f^+(\mathbf{r}) + f^-(\mathbf{r})] \\ \Delta f(\mathbf{r}) &= (f^+(\mathbf{r}) - f^-(\mathbf{r})) \end{aligned} \tag{8}$$

$q_x(N)$ represents the atomic charges associated with atom $x$ within the $N$-electron species. Recently, it has been shown[18h,i] that it is possible to use a QTAIM condensation scheme for Frontier Molecular Orbitals Fukui functions using finite differences within a topological partition. Such an approach has been shown to be particularly stable, having some advantage on other atomic evaluations of the Fukui functions scheme:

$$\sum_x f_x^\alpha = \sum_x \int_x |\varphi^{h(l)}(\mathbf{r})|^2 \, d\mathbf{r} = \int_x |\varphi^{h(l)}(\mathbf{r})|^2 \, d\mathbf{r} = \int_x f^\alpha \, d\mathbf{r} = 1 \tag{9}$$

where $h$ denotes HOMO, and $l$, LUMO, and the subscript $x$ under the integration sign indicates that the integration has to be performed only within the particular atomic domain of atom $x$.

As previouslsy demonstrated for the computations of electrostatic moments, any QTAIM local property computations can be performed using a topological ELF analysis. In this contribution, following the studies by Fuenteabla et al.,[18j] we present a Fukui analysis performed at both QTAIM and ELF levels.

**C. Computational Procedures.** The geometries of all formate−cation complexes were optimized using the hybrid functional B3LYP[19a,b] with the Jaguar 5.5 software.[20] The choice of the B3LYP functional was motivated by its observed good performance in the modeling of biomolecules containing Ca(II) and Mg(II) cations compared to MP2.[19b,c] We report in Supporting Information S1 some comparisons between different functionals, MP2 and CCSD(T), on optimal Ca(II) (or Mg(II))−formate geometries, confirming these findings. The LACV3P**[21] basis set combining a pseudopotential for the cation, and the all-electron 6-311G** basis set for the other atoms was employed. All geometries obtained with this less accurate energy function were then optimized further using the hybrid functional B3LYP but applying the all-electron 6-311++G** basis set[22] to all atoms, as provided by the Gaussian 03 software.[23] For the coagulation factors studied in the second part of this paper, single point calculations were employed for the two optimized malonate−cation complexes at the B3LYP/6-311++G** level of theory. All topological analyses were carried out using ELF grids of size 180 × 180 × 180 for moment analysis (300 × 300 × 300 for pictures) with the last version of the TopMoD90[24] package coupled to the TopChem[17] program providing DEMEP analysis. To compute the total molecular dipole, we have assumed as "global (or molecular) frame" the standard orientation provided by Gaussian 03, which computes molecular dipoles at the center of nuclear charges. B3LYP/CSOV computations were performed with the same basis set using an in-house version of HONDO 95.3,[25] whereas the GAMESS[26] software provided the RVS results computed at the Hartree−Fock level.

## Results

**A. Theoretical Description of Hard and Soft Metal-Cation Interactions with Carboxylate Moieties.** For blood coagulation proteins, X-ray studies clearly show that the main interactions involved in the biological activity of such enzymes involve networks built on the interaction of calcium or magnesium cations with carboxylate groups.[27] More precisely, X-rays unravel direct malonate−Ca(II)/or Mg(II) interactions.

To start our quantum chemical description of such proteins, we present here results on the interactions of differents metal cations with formate which are simple malonate models. As both monodentate and bidendate formate−cation coordinations are found in the structures, we have investigated the two cases. In a second part, we will focus on the specific interaction of calcium and magnesium cations with more realistic models directly extracted from the PDB structures of the different available factors.

*1. Topological Study of Hard and Soft Metal Cations: The Subvalence Concept.* In order to study the differences between metal cations, we have performed an ELF analysis upon DFT computations on several metal-cation−formate complexes encompassing monovalent cations such as Li(I), Na(I), K(I), and Cu(I) and divalent cations, namely, Mg(II), Ca(II), and Zn(II). They are displayed in Figure 1a and b. Indeed, in a recent study,[28] we showed that the density of Zn(II) exhibits a striking plasticity. However, Zn(II) binding ligands were shown to be able to adapt/redistribute their density according to their nature: sulfur atoms were shown to be the softest, being able to spatially delocalize their lone pairs, as oxygen and nitrogen mainly contract their lone pair volumes. The present study intends to generalize such observations. Indeed, striking differences can be observed by visually analyzing the obtained ELF topological pictures that can be associated with the well-known Parr and Pearson hardness concept linked to the resistance of an atom to change or deformity. One can see (Figure 1a and b) that the expected "hard cations" (high value of the $\eta_A$ hardness parameter, see Table 3 of ref 29) such as Li(I), Na(I), or Mg(II) have a spatial localization of their electron density condensed around the nucleus position. "Soft" cations, usually associated with lower $\eta_A$ values, exhibit specific splits within their outer-shell densities.

It is then possible to use *the concept of a "subvalence"* associated with outer-shell core basins which can be seen as the topological signature for a given hard or soft behavior of the metal. A quick look at the observed topological structures shows that the observed subvalent ELF basins are

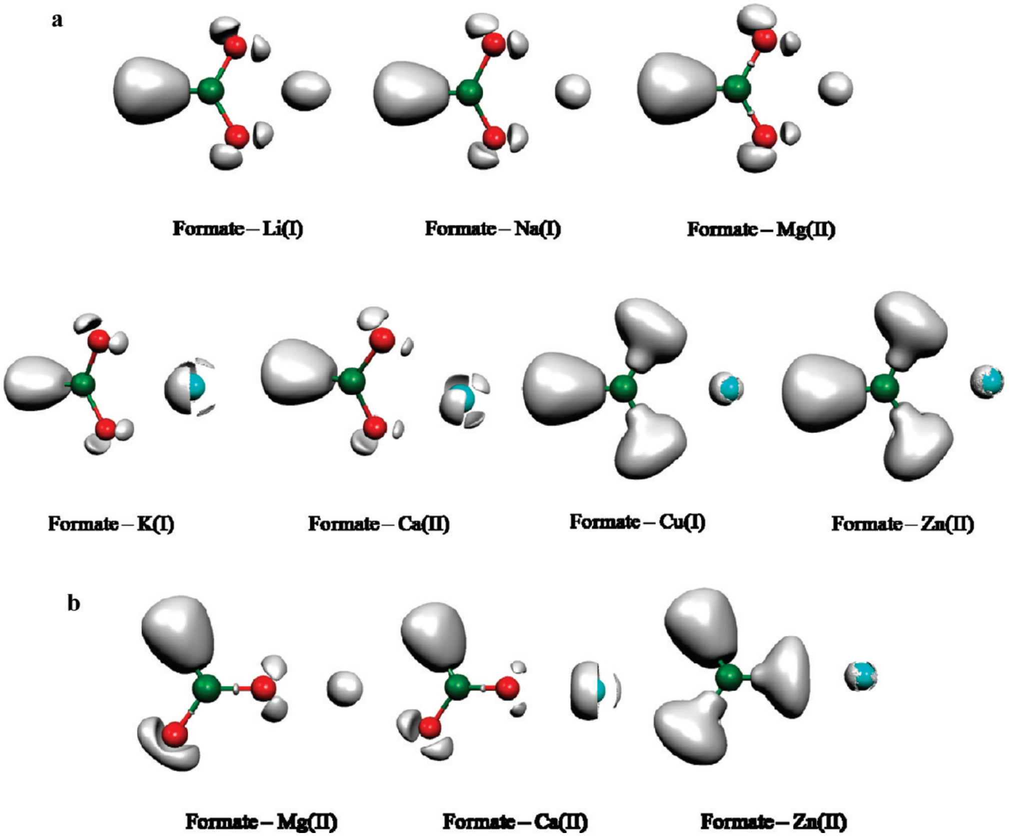


***Figure 1.*** (a) ELF representation of formate interacting in a bidentate mode with metal cations. Topological analysis of interactions between a formate and seven metal cations revealed, at the isosurface, a coefficient of 0.87 (except for Cu(I) and Zn(II) complexes, where it is 0.77). As seen in these pictures, electron densities remain condensed around the nucleus position for hard cations such as Li(I), Na(I), and Mg(II), whereas electron densities are split in four distinct subunits (basins), avoiding oxygen lone pairs, for softer cations such as K(I) and Ca(II), and split in two distinct subunits (basins), one of which is inserted between oxygen lone pairs and the cation, for soft cations such as Cu(I) and Zn(II). In soft cation complexes, a blue sphere describes the core electrons of the cations. In hard cation complexes, the spherical subvalence obscures the core electrons so that the blue sphere is not visible. (b) ELF representation of formate interacting in a monodentate mode with metal cations. Isosurface coefficients used to make theses pictures are the same as the ones used for part a. A very similar pattern is observed as for the bidentate formate−cation complexes. The Mg(II) electron density stays spherical. In the Ca(II) complex, three basins have merged into an annular one still avoiding oxygen lone pairs. The same split as for the bidentate complex is shown for Zn(II).

more numerous for soft cations, reflecting a more covalent character of their bonding to formate anions (see Figure 1).

*2. Subvalence: Understanding Physics at Play.* To have a deeper understanding of the physics taking place in such interactions, it is possible to extend the ELF analysis to the computation of local electrostatic moments (see Table 1) and Fukui functions (see Table 3) and to correlate them to RVS energy decompositions (see Table 2).

From these tables, we see that the more covalent the bonding character of the formate−cation intermolecular bond is, the greater the RVS cation polarization energy and local ELF cation dipole moment are. For example, a hard cation such as Na(I), which is poorly polarizable (weak polarizability), is involved in interactions dominated by electrostatics (Table 2) and does not show any split within its subvalence, whereas cations exhibiting stronger polarization and charge transfer interactions possess a higher number of basins. Moreover, for a covalently bonded very soft cation such as Zn(II), a subvalent Zn−O basin is observed between the formate oxygen lone pair and the metal, a hint of electron sharing. Ca(II), which is less soft and less covalently bonded, still exhibits a split, but subvalent basins remain distributed around the nucleus (Figure 1) and do not form any bond. Correlated CSOV energy decomposition computations have also been performed and are fully in line with HF RVS results. Details can be found in the Supporting Information, S2.

Observations of both monodentate and bidentate coordination modes show that "soft cation" subvalent basins clearly have the ability to orient themselves toward the formate oxygen lone pairs. That way, depending on its electron structure, each cation shows a specific topological signature which enables one to predict specific abilities of the cation to interact with its immediate environment thanks to the plasticity of its valence electron spatial organization. An indirect measurement of the soft/hard nature of the cations can be appraised by studying the volumes and density values of the formate oxygen lone pairs when interacting with cations. Hard cations such as Li(I) or Mg(II) clearly act on the lone pair densities which appear lower when compared to softer cations. Figure 2 exhibits the four oxygen lone pairs as they are when no metal cation interacts with the formate. Volume and density values reveal a dissymmetry between the internal and the external lone pairs, internal ones being

***Table 1.*** QTAIM and ELF $M_1$ and Dipole Moments of Formate−Cation Complexes[a]

| | | $M_1$ | | | $\mu$(D) | | |
|---|---|---|---|---|---|---|---|
| | | QTAIM | ELF | | −QTAIM | ELF | |
| | | Ω(M) | C(M) | V(M) | | | ab initio |
| formate | | | | | 0.92 | 0.90 | |
| Li | | | | | | | |
| | mono | 0.01 | 0 | | 9.41 | 9.40 | 9.39 |
| | bi | 0 | 0 | | 3.95 | 3.95 | 3.96 |
| Na | | | | | | | |
| | mono | 0 | 0.07 | | 11.60 | 11.67 | 11.69 |
| | bi | 0 | 0.08 | | 6.03 | 6.06 | 6.07 |
| K | | | | | | | |
| | mono | 0.21 | 0.31 | | 12.50 | 12.60 | 12.70 |
| | bi | 0.20 | 0.31 | | 7.20 | 7.30 | 7.30 |
| Mg | | | | | | | |
| | mono | 0.15 | 0.06 | 0.19 | 13.30 | 13.30 | 13.30 |
| | bi | 0.02 | 0.06 | | 7.30 | 7.30 | 7.30 |
| Ca | | | | | | | |
| | mono | 0.19 | 0.37 | | 14.80 | 14.90 | 14.90 |
| | bi | 0.16 | 0.35 | | 8.06 | 8.10 | 8.11 |
| Cu | | | | | | | |
| | mono | 0.18 | 0.47 | | 8.08 | 8.10 | 8.13 |
| | bi | 0.23 | 0.48 | | 2.97 | 3.01 | 3.04 |
| Zn | | | | | | | |
| | mono | 0.23 | 0.36 | 0.45 | 6.41 | 6.44 | 6.45 |
| | bi | 0.04 | 0.39 | | 3.06 | 3.09 | 3.10 |

[a] Values of cation's $M_1$ and dipole moment ($\mu$ expressed in Debye) for formate−cation complexes in both modentate (mono) and bidentate (bi) binding modes. $M_1$ is the polarization component of the total dipole moment (see text and Supporting Information). Concerning a cation, it is computed as the gap to the sphericity: the more a cation exhibits a spherical subvalence, the less its polarization and $M_1$ value are and vice versa. Two sets of M1 are reported: QTAIM values where all the electrons are gathered around the nucleus Ω(M) and ELF values where electrons are spread over the core basin C(M) and the subvalence basin V(M). QTAIM and ELF values for the total dipole moment are also reported; ab-initio dipole moments computed with Gaussian G03 software are given for comparison.

less populated and more contracted than the external ones. This is due to the fact that the internal lone pairs interact with each other because of the shorter distance between them. From Figures 1 and 2, it is possible to appraise the electronic redistribution within the oxygen lone pairs according to the presence or not of a binding metal cation and to its hardness or softness. Overall, it is important to point out that trends are conserved between ELF observations and Parr's hardness concept.[29] However, ELF pictures the final state of the cation electronic structure within the complex after cation−ligand orbital mixing and metal density relaxation (therefore, a feature also linked to its polarizability).

Concerning the specific Ca(II)/Mg(II) differences, our results demonstrate that overall less flexibility occurs in Mg(II) density compared to Ca(II), which tends to adjust to its immediate ligands. However, for Mg(II) in the monodentate binding mode, a slight increase of cation polarization associated with a topological split of its outer-shell density is noted, and this leads to the fact that Mg(II) could act slightly differently from usual hard cations. As we will see, this will have some consequences. Figure 3 shows a ELF representation of both monodentate and bidentate formate−Mg(II) complexes. A well-separated additional basin is found in the monodentate complex, where a partial charge is transferred. It worth noting that the Sr(II) cation which is sometimes found to substitute calcium under certain conditions (ref 30 and references therein) exhibits the same topological pattern as Ca(II).

To conclude, as the concepts of softness and hardness are involved, it is of importance to also consider the possibility of computing other popular local reactivity indicators, usually utilized to rationalize such phenomena. That way, we propose here an evaluation of the different "local" Fukui functions at both QTAIM and ELF levels. First, we computed such functions on an isolated formate molecule (see Table 3a and b). Again, the ELF Fukui analysis clearly shows the nonequivalence of the formate

***Table 2.*** RVS Energy Components for Selected Formate−Metal Cation Complexes[a]

| kcal/mol | | elec. | exch. | E1 | Epol | Epol(cation) | ECT | E2 | Etot |
|---|---|---|---|---|---|---|---|---|---|
| formate Li(I) | | | | | | | | | |
| | mono | −158.3 | 27.8 | **−130.4** | −17.6 | **−0.2** | −4.5 | **−22.1** | −152.6 |
| | bi | −179.2 | 28.1 | **−151.1** | −14.5 | **−0.1** | −6.7 | **−21.2** | −172.3 |
| formate Na(I) | | | | | | | | | |
| | mono | −138.0 | 20.9 | **−117.1** | −10.1 | **−0.3** | −0.2 | **−10.3** | −127.4 |
| | bi | −159.2 | 22.7 | **−136.5** | −8.3 | **−0.2** | −1.0 | **−9.3** | −145.8 |
| formate K(I) | | | | | | | | | |
| | mono | −125.9 | 26.4 | **−99.5** | −8.7 | **−1.8** | −1.3 | **−9.9** | −109.5 |
| | bi | −147.0 | 30.2 | **−116.8** | −7.1 | **−1.7** | −1.6 | **−8.6** | −125.4 |
| formate Mg(II) | | | | | | | | | |
| | mono | −300.0 | 43.7 | **−256.3** | −54.4 | **−0.4** | −7.8 | **−62.1** | −318.4 |
| | bi | −354.7 | 52.7 | **−302.0** | −48.4 | **−0.3** | −15.8 | **−64.1** | −366.1 |
| formate Ca(II) | | | | | | | | | |
| | mono | −287.7 | 76.3 | **−211.5** | −39.8 | **−2.3** | −17.8 | **−57.7** | −269.1 |
| | bi | −335.9 | 82.4 | **−253.5** | −32.8 | **−2.0** | −18.4 | **−51.2** | −304.7 |
| formate Cu(I) | | | | | | | | | |
| | mono | −177.6 | 68.1 | **−109.5** | −29.3 | **−15.3** | 8.4 | **−20.9** | −130.4 |
| | bi | −186.9 | 50.7 | **−136.2** | −19.1 | **−8.1** | 3.3 | **−15.8** | −152.1 |
| formate Zn(I) | | | | | | | | | |
| | mono | −319.5 | 66.7 | **−252.8** | −64.9 | **−4.5** | −6.7 | **−71.6** | −324.4 |
| | bi | −371.0 | 72.1 | **−298.9** | −56.0 | **−3.4** | −18.8 | **−74.8** | −373.7 |

[a] Values are given for the two monodentate (mono) and bidentate (bi) cation binding modes. elec. is the Coulomb electrostatic energy; exch. is the exchange repulsion. The sum of the two constitutes the first-order term E1. Epol and ECT are the polarization and charge transfer components of the second-order term E2, Etot being the sum of E1 and E2. An RVS decomposition energy allows us to separate the second-order terms over the constitutive fragments of a system. The individual polarization of each cation is reported. The more spherical the subvalence of a cation is, the less Epol is. The same pattern is observed for the ECT of each complex. It can be seen that ECT for the monodentate formate−Mg(II) complex is double that of the bidentate formate−Mg(II) complex. This is to be put in context with the appearance of the additional subvalence basin in the monodentate formate−Mg(II) complex (Figure 3).

***Table 3.*** (a) ELF Integrated Fukui Functions and Dual Descriptor Values for an Isolated Formate Molecule, (b) QTAIM Integrated Fukui Functions and Dual Descriptor Values for an Isolated Formate Molecule, (c) ELF and QTAIM Fukui Functions and Dual Descriptor Values for Selected Metal Cations[a]

|  | part a |  |  |  |
|---|---|---|---|---|
| formate basin | $f^-$(ELF) | $f^+$(ELF) | $f^0$(ELF) | $\Delta f$(ELF) |
| C(C) | 0.00 | 0.00 | 0.00 | 0.00 |
| C(O1) | 0.02 | 0.00 | 0.01 | 0.02 |
| C(O2) | 0.02 | 0.00 | 0.01 | 0.02 |
| V(C,H) | 0.22 | 0.30 | 0.26 | −0.08 |
| V(C,O2) | 0.04 | 0.01 | 0.02 | 0.03 |
| V(C,O1) | 0.04 | 0.01 | 0.02 | 0.03 |
| V(O1) | 0.20 | 0.01 | 0.10 | 0.19 |
| V(O1) | 0.13 | 0.03 | 0.08 | 0.10 |
| V(O2) | 0.14 | 0.03 | 0.09 | 0.11 |
| V(O2) | 0.20 | 0.01 | 0.10 | 0.19 |

|  | part b |  |  |  |
|---|---|---|---|---|
| formate atom | $f^-$(QTAIM) | $f^+$(QTAIM) | $f^0$(QTAIM) | $\Delta f$(QTAIM) |
| H | 0.17 | 0.15 | 0.16 | 0.02 |
| C | 0.05 | 0.04 | 0.05 | 0.02 |
| $O_1$ | 0.38 | 0.03 | 0.21 | 0.35 |
| $O_2$ | 0.38 | 0.03 | 0.21 | 0.35 |

|  | part c |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| cations | $f^-$(QTAIM) | $f^+$(QTAIM) | $f^0$(QTAIM) | $\Delta f$(QTAIM), isolated | $\Delta f$(ELF), isolated | $\Delta f$(QTAIM), complexed | $\Delta f$(ELF), complexed |
| Li | 1.00 | 0.23 | 0.61 | 0.77 | 0.70 | −0.14 | −0.11 |
| Na | 1.00 | 0.44 | 0.72 | 0.56 | 0.45 | −0.34 | −0.37 |
| K | 1.00 | 0.46 | 0.73 | 0.54 | 0.43 | −0.30 | −0.36 |
| Mg | 1.00 | 0.59 | 0.79 | 0.41 | 0.32 | −0.67 | −0.24 |
| Ca | 1.00 | 0.98 | 0.99 | 0.02 | 0.00 | −0.55 | −0.59 |
| Zn | 1.00 | 0.91 | 0.96 | 0.09 | 0.04 | −0.69 | −0.50 |

[a] Values are given for the isolated cations and for a cation within a bidentate formate−metal complex.

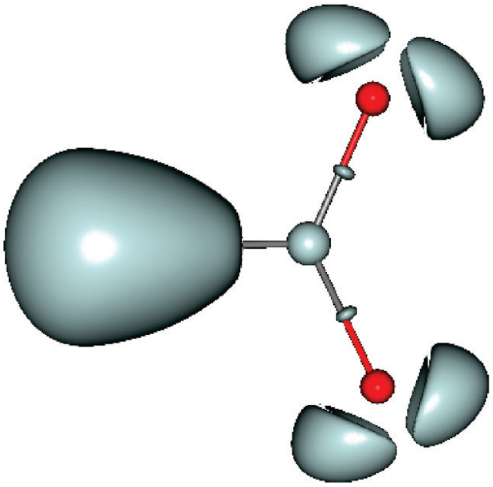


***Figure 2.*** ELF representation of a formate without a binding cation. This picture presents a formate in an uncomplexed state, where densities are not rearranged through interactions with a cation. Volumes of the oxygen lone pairs can be compared to the ones of formate−cation complexes shown in Figure 1a and b.

oxygen lone pairs, the external basins having different indicators from internal lone pairs. If QTAIM tends to show a uniform Fukui descriptor (the dual descriptor $\Delta f$ is always positive), ELF does not, as it provides a $\Delta f$ negative value on the C−H bond, providing insight about a possible different reactivity.

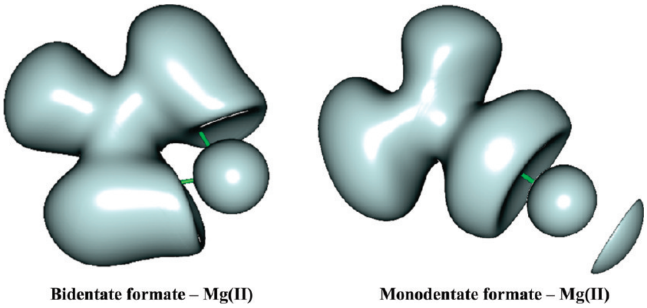


***Figure 3.*** ELF representation of formate−Mg(II) complexes at an isosurface coefficient of 0.22. The electron density of the two formate−Mg(II) complexes are compared in this picture. Contrary to the bidentate formate−Mg(II), where the subvalence stays spherical, in the monodentate formate−Mg(II), an additional basin appears at the opposite side of the coordination to the oxygen. This is consistent with the augmentation of both $M_1$ shown in Table 1 and charge transfer energy shown in Table 2 for this complex.

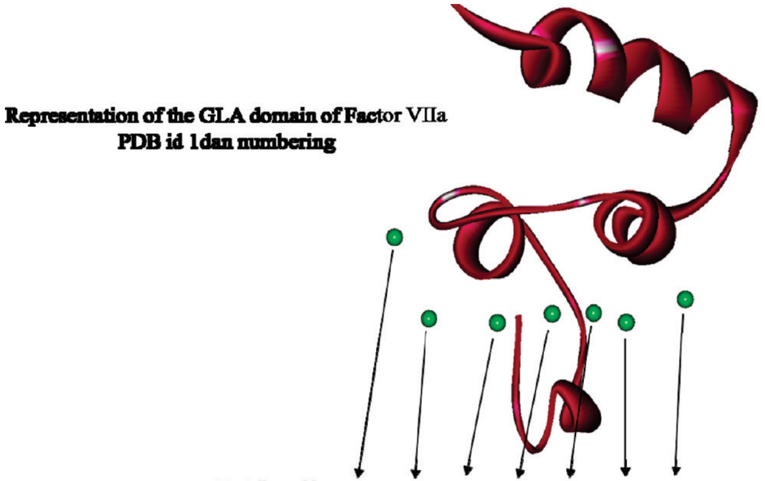
Concerning the metal cations, Table 3c brings interesting information. Let us first consider first the isolated metal cations. If all cations exhibit $f^-$ values of 1, they have very different $f^+$ values. Again, following chemical intuition, strong differences occur between hard and soft cations. Hard cations such as Li(I) exhibit low values of $f^+$, whereas soft cations such as Zn(II) have $f^+$ values tending toward 1. The dual descriptor appears then to be a good global indicator reflecting the cation's ranking as $\Delta f$ tends to 0 with increasing cation softness. Table 3c depicts the results for selected bidentate formate cations. Global trends are preserved despite strong, different bonding modes. Again, the softer cations exhibit smaller $\Delta f$ values (here more negative) than harder ones. If they agree well, one difference between ELF and QTAIM values is observed: ELF tend to show more difference between Ca(II) and Mg(II) than QTAIM. This is reflected by a difference of 0.2 (QTAIM) vs 0.3 (ELF) in the $\Delta f$ values between the two cations.

Of course, the numerical values strongly depend on the (never unique) topological partition scheme and therefore on the density attribution to atoms/centers,[18h] but the ELF and AIM approaches are not subject to strong basis set/diffuse function dependence[18a,h] and are quite stable. Overall, beyond the numbers, an interesting qualitative agreement is observed and supports the previously depicted ELF subvalence basins and the numerical values extracted from other interpretative techniques described above.

**B. Metal Cation's Electron Structure/Biological Activity Relationship in Coagulation Proteins.** We use here the commonly accepted terminology, first, to differentiate the $\gamma$-carboxyglutamic acid itself, identified as Gla, from the $\gamma$-carboxyglutamic acid-rich domain, identified as GLA, second, to name the first 11 residues at the N-terminal extremity of the $\omega$-loop, and third to specifically identify residues 4, 5, and 8 (5, 6, and 9 for FIX) from the previous sequence as the keel. In addition, since we have to compare several GLA domain crystal structures, each having their Ca(II) or Mg(II) positions differently numbered, we will use for clarity purposes the same numbering for all, that is, the one found in the crystal structure of the human factor VII (1DAN, see ref 38).

*1. Metal Cation's Electron Structure/Biological Activity Relationship in Coagulation Proteins.* When a blood vessel is injured, a cascade of protein−protein interactions[1] rapidly occurs, leading to the formation of a cross-linked fibrin clot, eventually restoring the integrity of the circulatory system. A number of proteins involved at the early stage of the process are vitamin-K-dependent zymogens of serine proteases.[31] Among this family of coagulation factors, factor VII, once activated and bound to tissue factor (TF), activates zymogen factors IX and X into functional factors IXa and Xa. These in turn transform prothrombin into active thrombin, which is ultimately responsible for the conversion of fibrinogen to fibrin.[32] A cell-based model of coagulation, which incorporates the important roles of endothelial cells and platelets, has recently been introduced, and this more physiological view appears to be gaining acceptance.[33] Many other factors and cofactors are necessary for the completion of the mechanism, but we will focus only on two steps.[1]

*2. Structural Analysis and Biological Activity.* The above-mentioned three factors are made up by four domains. At the N extremity, we find a GLA domain, followed by two Epidermal Growth Factor (EGF) like domains, namely, EGF1 and EGF2, terminated by a Serine Protease (SP) domain. The primary structures of GLA domains (residues 1 to 48) are highly conserved:[34] only a few residues differ along the sequences of factors VII, IX, and X. Remarkably, the first nine Gla residues, as well as two cysteines, are always found precisely at the same place. Accordingly, for all factors, two hydrophobic residues are adjacent to the first group of two Gla residues: Phe4 and Leu5 for FVIIa and FXa factors and Leu6 and Phe9 for FIXa. We can also notice the presence of an Asn residue in the second position of the sequence. Moreover, analysis of the PDB files for these factors shows that the secondary structures are also very similar and essentially superimposable. From the N extremity to the end of the domain, we find successively the $\omega$-loop with its two above-mentioned hydrophobic amino acids along with either another hydrophobic residue (FVIIa−Leu8 and FXa−Met8) or a cationic one (FIXa−Lys6) pointing their side chains toward the exterior of the protein. Three α-helices are also found. The first two are parallel and maintained so by a disulfide bridge established by the two cysteines cited before. All Gla residues are distributed along these three subunits: Gla6 and -7 are at the top of the central loop of the $\omega$-loop; Gla14, -16, -19, and 20 belong to the first α helix, whereas Gla25, -26, and -29 belong to the second. The third α-helix links the GLA domain to the remaining part (3rd helix) of the protein. A Gla residue can also be found in this part of the domain. In all examined X-ray structures, eight cations are present.[38] One cation is located at the hinge between the third α helix and the first chain of the EGF1 domain. The other seven cations are found aligned at the interstice between the base of the two parallel α helices and the top of the $\omega$-loop. In this zone, two groups of cation binding sites can be defined. The first group is called "high affinity" Ca(II) binding sites[39] and is constituted by the 5 inner Ca(II) sites (numbered 3, 4, 5, 6, and 8, respectively).[38] Within these sites, Ca(II) were found coordinated 6, 7, 7, 7, and 3 times, respectively, at distances in the 2.4−2.8 Å range in the very first GLA domain ever structurally determined (1992), namely, the GLA domain of Ca−Prothrombin Fragment I (see Figure 4 and Table 3 of ref 35). Subsequently, several authors[36,37,41] reported very similar coordination numbers and distances for the factors studied here. The second group, called "low affinity" Ca(II) binding sites, defines the two external sites. At these positions, cations can be either Ca(II) or Mg(II). Figure 4 lists the cation binding sites present in different PDB structures of the studied factors' GLA domains. Occupation by either Ca(II) or Mg(II) in all binding sites seems to be due to details in the preparation of the protein crystal during the X-ray crystallization process. In fact, F VIIa (1DAN),[38] F IXa (1J35),[39] and F Xa (1IOD)[40] were prepared using only $CaCl_2$ as crystallographic salt. It is noticeable that all the binding sites are occupied by calcium, solely present in the environment. When $MgCl_2$ is added to $CaCl_2$, at physiological concentration, for the preparation of F IXa (1J34),[39] all the inner binding sites are occupied by calcium, while the external binding sites are occupied by magnesium. But, prepared under the same conditions, a recent X-ray structure deposited in the Protein Data Bank by Bajaj et al. shows Mg(II) residing in the central site number 5 replacing Ca(II) in the GLA domain of F VIIa (2A2Q).[41] However, when only $MgCl_2$ is used for the preparation of F Xa (1P0S),[4] the two external sites contain a Mg(II) cation, but the sole central site number 5 shows a third Mg(II) cation. In addition, contrary to the five other



| Coagulation Factor | PDB id | Crystallographic Salt | Site 7 | Site 8 | Site 6 | Site 5 | Site 4 | Site 3 | Site 9 | ω-Loop Present ? | Resolution (Å) | Ref. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| F VIIa | 1dan | $CaCl_2$ | Ca(II) | Ca(II) | Ca(II) | Ca(II) | Ca(II) | Ca(II) | Ca(II) | YES | 2.00 | [26] |
| F VIIa | 2a2q | $CaCl_2+MgCl_2$ | Mg(II) | Ca(II) | Ca(II) | Mg(II) | Ca(II) | Ca(II) | Mg(II) | YES | 1.80 | [30] |
| F IXa | 1j34 | $CaCl_2+MgCl_2$ | Mg(II) | Ca(II) | Ca(II) | Ca(II) | Ca(II) | Ca(II) | Mg(II) | YES | 1.55 | [27] |
| F IXa | 1j35 | $CaCl_2$ | Ca(II) | Ca(II) | Ca(II) | Ca(II) | Ca(II) | Ca(II) | Ca(II) | YES | 1.80 | [27] |
| F Xa | 1iod | $CaCl_2$ | Ca(II) | Ca(II) | Ca(II) | Ca(II) | Ca(II) | Ca(II) | Ca(II) | YES | 2.30 | [28] |
| F Xa | 1p0s | $MgCl_2$ | Mg(II) | | | Mg(II) | | | Mg(II) | NO | 2.80 | [29] |

***Figure 4.*** Description of metal cation coordination sites of six GLA domains. The picture is a representation of the GLA domain of F VIIa (1DAN), exhibiting its constitutive parts including the seven cation binding sites. For each site, a comparison of six X-ray geometries shows its occupation by either a Mg(II) or a Ca(II) cation, according to the crystallographic salt used for the preparation.

structures that contain it, no $\omega$-loop is present in the structure of this domain, its constitutive peptide seeming to "float" in the environment. From these observations, it may be deducted that Ca(II) cations are necessary in the central zone to structure the $\omega$-loop and that the external sites are usually occupied by magnesium. Close examination of the binding of the $\omega$-loop to the rest of the GLA domain reveals the network of interactions between the amino acids borne by the $\omega$-loop, the cations and the amino acids present in the two antiparallel α helices. In all factors where the $\omega$-loop is present, except for F VIIa with Mg(II) in the central site, the $NH_3^+$ extremity of Ala1 (or Tyr1 for F IXa) establishes three H bonds with the surrounding residues, namely, carbonyl O of Gln21 (F VIIa), Ala21 (F Xa), or Lys22 (F IXa); $O\epsilon_4$ of Gla20 (F VIIa and F Xa) or Gla21 (F IXa); and $O\epsilon_4$ of Gla26 (F VIIa and F Xa) or Gla27 (F IXa). In F VIIa (2A2Q), however, a single H-bond remains with $O\epsilon_2$ of Gla26. A comparison between coordination numbers and coordination distances for each cation of the two structures of F VIIa (1DAN vs 2A2Q), as given in Table 2 of ref 41 by Bajaj et al., reveals that, when only Ca(II) is present in the five inner sites, direct coordinations of cations by the carbonyl O of Ala1 (or Tyr1 for F IXa); $O\delta_1$ of Asn2; $O\epsilon_1$ and $O\epsilon_4$ of Gla6; and $O\epsilon_1$, $O\epsilon_2$, and $O\epsilon_4$ of Gla7 are observed. By contrast, in F VIIa (2A2Q), where the central Ca(II) is substituted by Mg(II) with two water molecules (S209 and S363) completing the coordination sphere of this cation, the same interactions are observed, but this time through a network of eight water molecules (S209, S262, S363, S411, S508, S602, S694, and S722), present in between the cations and the $\omega$-loop. As a consequence, the $\omega$-loop has moved down approximately 0.5 Å in order to leave enough space for water to insert (see Figure 2A of ref 41).

The existence of the $\omega$-loop is of much importance in that the very first step of the coagulation process is for coagulation factors to colocalize on cell surfaces.[42] In that step, the biological function of the GLA domain is first directly responsible for the linkage of the factors to the membranes, and ultimately to the fibrin clot. This can be done thanks to the three hydrophobic residues of the $\omega$-loop[43] (or two hydrophobic residues and a cationic one in F IXa), oriented in such a way that they are able to dive deeply inside the membrane and interact through hydrophobic bonds with neighboring lipids,[43,44] in addition probably through a buried salt bridge involving Lys5 of F IXa.[43] The deep insertion of the GLA domain inside the membrane also allows the creation of an interaction between a phosphatidylserine (PS) of the membrane and the Ca(II) cation present in binding site number 8, strengthening the anchorage of the coagulation factor inside the cellular membrane.[45] Inhibition of the GLA domain by direct ligand bonding to the $\omega$-loop (such as snake venom protein, see ref 39) is responsible for the loss of membrane linking with a subsequent loss of the coagulation process.

*3. Theoretical Study of Interactions of Ca(II) vs Mg(II) with Malonate Groups.* In this section, we systematically supplement the ELF analysis with detailed RVS energy decomposition results.

In GLA domains, the transformation of glutamic acids in Gla is realized by the action of vitamin K and several specific enzymes with the addition of a carboxylate group to the $\gamma$-carbon of the glutamate;[46] two carboxylate groups borne by the same $\gamma$ carbon constitutes a malonate group. Two malonates coordinating a metal cation is one of the unit structures observed in GLA domains, displaying as many as four monodentate formate−cation interactions. Two

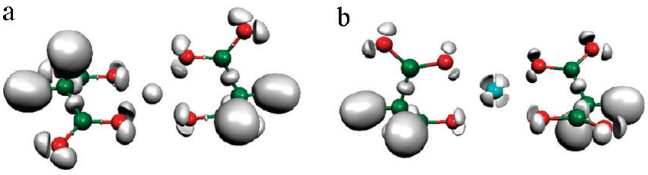


***Figure 5.*** (a) Optimized geometry of a two-malonate−Mg(II) complex. (b) Optimized geometry of a two-malonate−Ca(II) complex. All the ELF pictures were revealed at the 0.87 isosurface coefficient. These pictures exhibit the perfect tetrahedral binding mode of the cations. Mg(II) subvalence is spherical, and the Ca(II) one is split in four well separated basins recalling the bidentate formate−cation complexes.

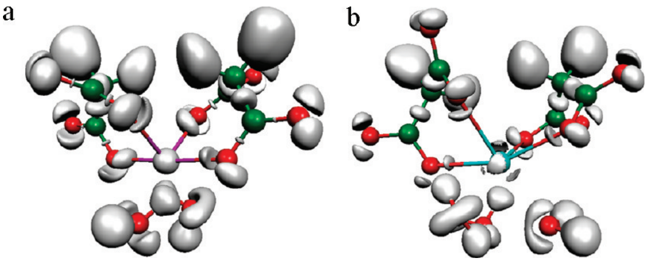


***Figure 6.*** (a) Extracted two-malonate−Mg(II) "deformed tetrahedral" complex. (b) Extracted two-malonate−Ca(II) "deformed tetrahedral" complex. The Mg(II) complex is binding site no. 7 of F IXa (1J34); the Ca(II) one is binding site no. 9 of F Xa (1IOD). Water molecules complete the coordination sphere of each cation: six for Mg(II) and seven for Ca(II). Cation coordinations are also disclosed in order to evidence the "deformed tetrahedral" binding mode of the cations. Each cation behavior does not change as the number of coordinations increases.

different computations at the same level of theory (B3LYP/6-311++G**) have been performed on these systems: a single point calculation using directly extracted geometries from PDB structures on which H atoms were added and geometry optimizations performed using the previous systems as starting points. At the end of the optimization process, the obtained complexes show the two approximate planes of malonates perpendicular to each other around the metal cation, itself four times tetrahedrally coordinated with coordination distances of approximately 2.1 Å for the magnesium and 2.4 Å for the calcium. A calcium complex was extracted from the PDB structure of F Xa (1IOD), whereas the magnesium complex was extracted from the PDB structure of F IXa (1J34). From these two systems, external binding sites (number 7 or 9) are selected, in which cations are surrounded by solvating water molecules: two for Mg(II) and three for Ca(II). Figures 5a and b show the ELF topological analysis on the two optimized geometries. As expected, Mg(II) (Figure 5a) does not exhibit any split of its subvalence, whereas Ca(II) (Figure 5b) has its subvalence split into four well separated basins, oriented in such a manner that no oxygen lone pair faces a cation basin. The same pattern as described above is observed in Figure 6a for Mg(II) and Figure 6b for Ca(II), within the extracted geometries. Thus, upon increasing the coordination number from two in the bidentate formate−cation complexes to four in the optimized geometries shown in Figures 5a and b and ultimately to their maximum in the extracted complexes, each cation consistently exhibits the same behavior. Table 4 reports the results of the RVS analysis, disclosing the individual contributions of the interaction energies for the two malonate−cation complexes shown in Figures 5a and b and 6a and b. These results show first a greater electrostatic term for Mg(II) less compensated by the repulsion term than for Ca(II). As a result, the excess of electrostatic energy reflected by the negative first order term of Mg(II) over Ca(II) is −90 kcal/mol for the optimized complexes and less than −77 kcal/mol for the extracted geometries, reflecting the hardness of the Mg(II) cation. Since a RVS analysis can separate the second order energies for each constitutive monomer of a many-body system, the individual polarization energies of the cations are also reported. It can thus be observed that, as Ca(II) moves away from its optimal geometry, its polarization increases, and that Mg(II) is essentially not polarized. The charge transfer contribution in the Ca(II) complexes decreases upon moving away from the optimal geometry but remains significant, whereas it is null for the Mg(II) optimized complex and slightly increases in the

***Table 4.*** Theoretical RVS Analysis of Tetrahedral Cation Binding Sites[a]

| kcal/mol | 2 malonates + Mg(II) | 2 malonates + Mg(II) | 2 malonates + Ca(II) | 2 malonates + Ca(II) |
|---|---|---|---|---|
| geometry | optimized | F IXa 1J34 | optimized | F Xa 1IOD |
| figure number | 5a | 6a | 5b | 6b |
| electrostatic | −777.7 | −703.3 | −709.0 | −636.8 |
| repulsion | 90.8 | 68.2 | 111.5 | 78.2 |
| first order energies ($e_1$) | −686.9 | −635.1 | −597.5 | −558.6 |
| **polarization of cations** | **−0.1** | **−0.3** | **−0.6** | **−2.2** |
| polarization | −94.2 | −90.5 | −63.0 | −61.5 |
| charge transfer | 0.0 | −1.4 | −11.2 | −9.7 |
| second order energies ($e_2$) | −94.2 | −91.9 | −74.2 | −71.2 |
| total $e_1 + e_2$ | −781.1 | −720.9 | −671.7 | −629.8 |

[a] RVS energy decomposition has been performed upon theoretical models of tetrahedral binding sites. They are built on the same number of atoms (no water molecules are considered) in order to directly compare the energy contributions. Optimized geometries are the extracted ones on which the optimization process was performed. Mg(II) complexes' first order terms are greater than the ones of Ca(II) complexes, whereas Ca(II) cations are more polarized and generate a larger charge transfer than for Mg(II). This is consistent with the Mg(II) cation having a spherical subvalence entering in more electrostatic types of interactions and the Ca(II) cation splitting its subvalence and creating more interactions of covalent type.

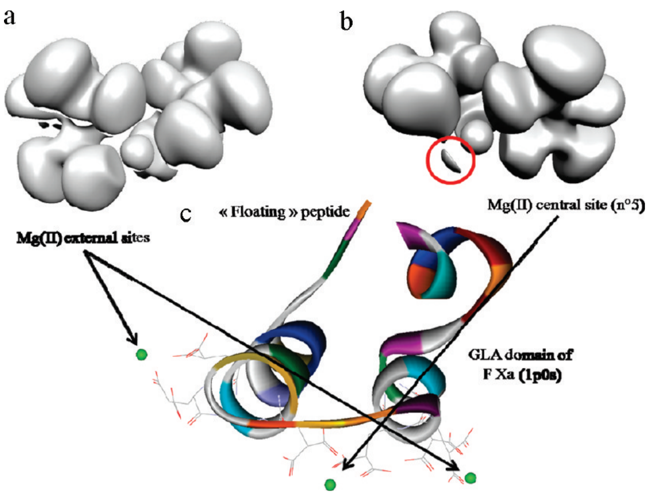


***Figure 7.*** (a) “Hard” Mg(II) binding mode. (b) “Soft” Mg(II) binding mode. (c) Localization of the two binding modes within F Xa (1P0S). Electron densities are revealed at the isosurface coefficient of 0.22. The additional basin shown in the red circle is confirmed by the increase of the charge transfer energy illustrated in Table 4 (−7.6 kcal/mol). As is the case when only Mg(II) is present in a vitamin-K-dependent coagulation factor environment, the cations occupy only the three tetrahedral binding sites, being unable to structure the $\omega$-loop, and leaving its constitutive peptide floating.

complex extracted from X-ray crystallography. It can be deduced that, contrary to Mg(II), Ca(II) is polarized as it is in the bidentate formate−cation type of complex and that the charge transfer contribution becomes selective from one cation to another in such complexes. Thus, Mg(II) enters a more electrostatic interaction, whereas a polarized Ca(II), generating a stronger charge transfer, enters a more covalent interaction. It is important to point out that we also perform such analysis in the presence of an additional PCM implicit solvent. We did not observe any changes in the topology of the system. Such results can be found in Supporting Information S3.

*4. Theoretical Study of the Selectivity Ca(II)/Mg(II). a. Is the Mg(II) Cation Strictly Hard in the Gas Phase?* In the GLA domain of F Xa (1P0S), in which only magnesium ions are present in the environment, a Mg(II) is observed to be three times coordinated to oxygen atoms of Gla 16 and 26, in the 2.3−2.5 range of distances. Figures 7a and b compare the topology of Mg(II) in two positions: the first one (Figure 7a) comes from an external cation site (site no. 7 of F IXa 1J34); the second one (Figure 7b) is the central site (no. 5). Due to the low resolution (2.8 Å) of the crystal structure of F Xa (1P0S), no water molecule is resolved near the cation in the central site no. 5. Despite the fact that water must be present within the coordination sphere of Mg(II) in a real enzyme, no water molecule is considered in the present theoretical gas phase study, so that the coordination number is four for the external site and three for the central site. As can be seen in Figure 7a, the cation does not exhibit any split of its subvalence, whereas in Figure 7b, the presence of the additional basin (red circle), recalling the monodentate formate−Mg(II) binding mode, clearly suggests that Mg(II) transfers a part of its density into this additional basin. This is confirmed by the magnitude of the RVS charge transfer contribution (see FXa 1P0S in Table 5), namely, −7.6 kcal/mol, instead of a null charge transfer in the optimized geometry. In this particular case, it is important to point out that Mg(II) does not have a hard cation behavior since it exhibits some subvalence capabilities. However, if such a finding is important for understanding the electronic distribution in Mg(II) complexes, could it have an impact when considering more realistic condensed phase GLA domains containing explicit solvent molecules?

*b. Selectivity of Ca(II) vs Mg(II) Cations within the Six Gla Domains.* From Figure 4, we have found three possible binding sites for Mg(II), namely, sites no. 7 and 9 (external) and 5 (central), the latter exhibiting either a Mg(II) or a Ca(II). The other sites are exclusively occupied by Ca(II) cations, a necessity to structure the $\omega$-loop in its functional geometry. These observations have been confirmed by several experiments, in which Mg(II) and Ca(II) were added at physiological concentration to a previously divalent cation-free environment.[2] When only Mg(II) is present, no enzymatic activity occurs. This is to be put in relation with the

***Table 5.*** RVS Analysis of Selected Cation Binding Sites, Extracted from X-Ray Structures[a]

| kcal/mol | 2 malonates + Mg(II) | 2 malonates + Mg(II) | Mg(II) hexa coordinated | 2 malonates + Ca(II) | Ca(II) octa coordinated |
|---|---|---|---|---|---|
| **Water molecules** | **2** | **0** | **3** | **3** | **2** |
| PDB ID | F IXa 1J34 | F Xa 1P0S | F VIIa 2A2Q | F Xa 1IOD | F IXa 1J34 |
| figure number | 6a | 7b | 8a | 6b | 8b |
| electrostatic | −696.7 | −577.9 | −746.5 | −690.6 | −598.5 |
| repulsion | 130.0 | 23.7 | 184.5 | 188.3 | 111.1 |
| first order energies ($e_1$) | −566.7 | −554.3 | −562.0 | −502.3 | −487.4 |
| **polarization of cations** | **−0.4** | **−0.4** | **−0.1** | **−1.8** | **−0.5** |
| polarization | −122.8 | −66.5 | −92.6 | −95.0 | −33.0 |
| charge transfer | −9.8 | −7.6 | −2.7 | −20.7 | −4.9 |
| second order energies ($e_2$) | −132.5 | −74.1 | −95.3 | −115.7 | −37.9 |
| total $e_1 + e_2$ | −699.2 | −628.4 | −657.3 | −618.0 | −525.3 |

[a] RVS energy decomposition has been performed upon various extracted geometries in order to confirm the theoretical results on realistic systems. Thus, all the components, including water molecules, were taken into consideration. The number of water molecules is detailed for each complex, as well as the X-ray structure from which the system was extracted and the corresponding figure referenced. The same trend is observed, as for theoretical tetrahedral systems, concerning electrostatic energies favoring Mg(II) complexes, and the Ca(II) cation being more polarized, producing a greater charge transfer.

F Xa (1P0S) structure where the $\omega$-loop is not formed. When only Ca(II) is present, the enzymatic activity is signicantly higher but can be enhanced by the addition of Mg(II). The three possible Mg(II) sites are constructed from three pairs of malonates provided by site no. 5 Gla16 and 26 (Gla17 and 27 for F IXa), site no. 7 Gla14 and 19 (Gla15 and 20 for FIXa), and site no. 9 Gla25 and 29 (Gla26 and 30 for F IXa). Close examination of these sites reveals that they all share the same pattern in which the cation is coordinated by two groups of two oxygens residing in the same side of two almost parallel malonate groups. Such geometry recalls the optimized geometry of the two malonates bound to a cation complex. Because they are not perfectly tetrahedral, these three binding sites will be called “deformed tetrahedral” cation binding sites. Figures 6a and b show ELF pictures of such a “deformed tetrahedral” binding site occupied by a Mg(II) cation as extracted from site no. 9 of F IXa (1J34) and a Ca(II) cation as extracted from site no. 7 of F Xa (1IOD), respectively. From Table 4, RVS total energies give the preferrence of Mg(II) over Ca(II) within this type of site by a difference of −109.4 kcal/mol for the optimized geometries and −91.1 for the extracted ones. This is due to the excess of first order energies in favor of Mg(II), namely, −89.4 and −76.5 kcal/mol. This is consistent with Mg(II) not splitting its subvalence and thus being involved in more electrostatic types of interactions. The presence of water molecules (two for Mg(II) and three for Ca(II)) within the coordination sphere of cations does not modify this conclusion (see Table 5).

The four other sites, however, are less structured, as their Ca(II) cations are coordinated four to eight times, in both monodentate and bidentate mode, by several carboxylate groups belonging to different Gla residues, as well as by carbonyl groups of the backbone of Ala1 (Tyr1 for F IXa) and the side chain of Asn2. For example, in site no. 6 of F IXa (1J34), Ca(II) is found octa coordinated, by Tyr1, Gla21 in a monodentate mode, Gla7 and -17 in a bidentate mode, and two water molecules (Figure 8b). The now familiar split of the subvalence of Ca(II) is observed. An ELF computation performed on this model reveals that the net charge of the cation within the system is +1.67 instead of +2 in the fundamental state. This means that up to one-third of an electron was transferred from the ligand oxygen lone pairs to the Ca(II) cation, with Ca(II) consistently able to adapt its electronic density to its environment by splitting its subvalence, thus entering into more covalent types of interactions. This capacity of adaptation for Ca(II) explains why, when only present in a coagulation factor environment, all the binding sites are loaded with Ca(II) (F VIIa 1DAN, F IXa 1J35, and F Xa 1IOD).

Therefore, with respect to the experimental results shown in Figure 5 of ref 2, a mechanism can be proposed for the activation of vitamin-K-dependent coagulation factors by Mg(II) and Ca(II) cations. Upon introducing first Mg(II) to a cation-free environment of an enzyme, one Mg(II) cation goes to each of the three “deformed tetrahedral” binding sites (nos. 5, 7, and 9), completing its coordination sphere with two water molecules. At this stage, the enzyme is not active and the $\omega$-loop is not structured (F Xa 1P0S). When Ca(II) is added, one Ca(II) cation goes to each of the four other sites (nos. 3, 4, 6, and 8). This anchors Ala1 (or Tyr1 for F IXa) and Asn2 to the rest of the GLA domain[47] and folds the peptide in this particular $\omega$ geometry through interactions involving Gla6 and -7 (Gla7 and -8 for F IXa) and the rest of the domain (F VIIa 2A2Q). However, prepared with the same mix of cations at physiological concentration[39] as the one used to prepare F VIIa (2A2Q)[41] for the crystallization process, F IXa (1J34) exhibits a Ca(II) cation in central site no. 5 instead of a Mg(II) cation. Could the occupation of the central binding site no. 5 by either a Mg(II) (2A2Q) or a Ca(II) (1J34) be explained by the topological difference of the two cations?

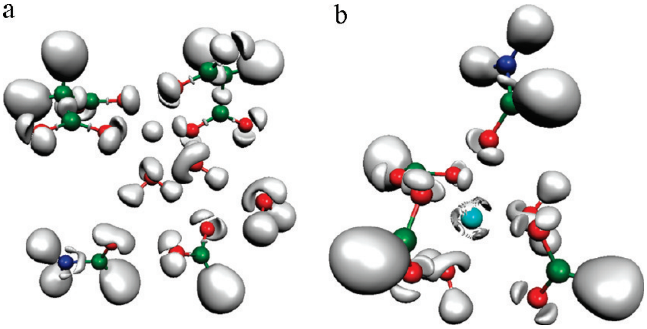


***Figure 8.*** (a) Extracted geometry of binding site no. 5 of F VIIa (2A2Q) with Mg(II). (b) Extracted geometry of binding site no. 6 of F IXa (1J34) with Ca(II). Part a shows a tetrahedral binding site loaded with a Mg(II) cation, that structure being the $\omega$-loop through a network of H-bonds established by highly polarized water molecules coordinated to the cation. Part b shows direct interactions between a Ca(II) cation and some constitutive segments of the $\omega$-loop. The subvalence pattern of the two cations stays consistent with the one observed in all the studied geometries, whatever the size of the system considered.

*c. Direct Interactions with Cations vs Interactions with Cations through Water.* We consider here a system constituted by the five central Ca(II) binding sites extracted from the geometry of the GLA domain of F VIIa (1DAN), which is very similar to the one of F IXa (1J34). This complex was built from the backbone of Ala1, a formamide group representing the side chain of Asn2, and two malonates for Gla6 and -7, with Ala1, Asn2, Gla6, and Gla7 being part of the $\omega$-loop, in addition to three malonate groups from Gla16, -20, and -26 and a formate given by Gla29 and the five Ca(II) ions. In addition, up to six water molecules were placed on the calcium ions according to their position in the X-ray structure. This system, which is globally neutral, is an X-ray crystal snapshot where hydrogens and waters have been added, focusing on the interactions between Ca(II) cations and their ligands. Such a system is found in X-ray crystal structures 1DAN, 1J34, 1J35, and 1IOD in which cations directly interact with their environment. Figure 9 exhibits the ELF study on this system. This picture reveals the network of interactions between the subvalence basins of the cations and the lone pairs of the coordinating oxygens. The split of all Ca(II) subvalences indicates that these interactions are partially covalent because they are built from electronic exchanges between the cations and the lone pairs present in their immediate environment. This results in a

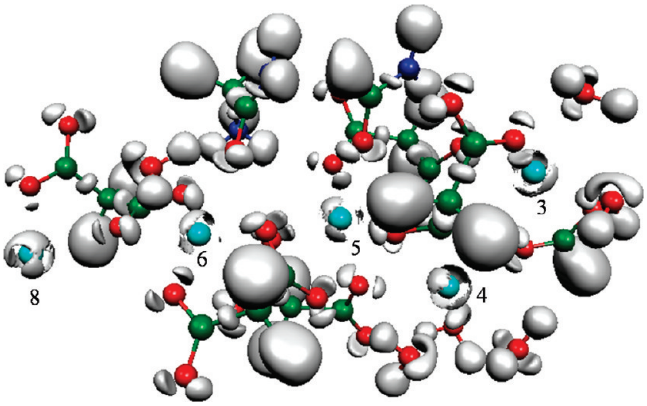


***Figure 9.*** Topological analysis of the five central Ca(II) binding sites of F VIIa (1DAN) GLA domain. This picture unravels the network of charge transfer interactions between the five Ca(II) cations and the oxygen lone pairs borne by Gla residues of the two upper α-helices, on one hand, and the oxygen lone pairs borne by Ala1 (backbone), Asn2 (side chain), and Gla6 and -7 of the lower *ω*-loop, on the other hand. These interactions structure the *ω*-loop in its functional geometry. It can be noticed that Ca(II) cations show different patterns of their subvalence split according to their different numbers of coordinaton. The numbering of the calcium ion binding sites of Figure 3 is used.

network of many-body interactions at the center of the GLA domain. QTAIM values of $M_1$ of each Ca(II), computed with the ELF function, confirm the intensity of the charge transfer. Ca(II) no. 3 and 5, each being 7-fold coordinated, have the lowest value of the series with a $M_1$ value of 0.041 D and a net charge of +1.66 instead of +2, whereas Ca(II) no. 4 and 6, with 6-fold coordination, exhibit enhanced values, namely, 0.060 D and +1.68, for $M_1$ and the charge, respectively. Ca(II) no. 8, being four times coordinated, shows values of $M_1$ very close (0.166 D, +1.72) to the one found for the small bidentate formate−cation complex (0.160 D and +1.70 for $M_1$ and the charge, respectively, see Table 1). This indicates that the greater the coordination number, the less the polarization and therefore the residual charge.

In the GLA domain of F VIIa (2A2Q), the central Mg(II) is found six times coordinated: four times tetrahedrally by malonates of Gla16 and -26 and two times by water molecules S209 and S363 (Figure 8a). The formamide moiety of Asn2, a formate group from Gla7, and a third water complete the network of H-bonds through which this central Mg(II) contributes to structure the *ω*-loop. As seen in Figure 8a, the central Mg(II) subvalence remains almost perfectly spherical, as in the bidentate formate−Mg(II) complex and in the two-malonate−Mg(II) tetrahedral optimized geometry. Close examination of the RVS polarization contribution (Table 5) reveals that, if the cation is as expected almost not polarized, the polarization of the three water molecules accounts for approximately 30% of the total polarization of −92.7 kcal/mol. Indeed, one of the water molecules is highly polarized and bears an Epol(RVS) value of −13.9 kcal/mol. Thus, interactions between the cation and the *ω*-loop are established through highly polarized water molecules.

While the presence of Ca(II) is required[2,3,48] in the central region for an effective coagulation function, it has been demonstrated that the affinity of coagulation factors for either cellular membranes,[45,49] tissue factors,[5] or anticoagulant agents[36,40] is strongly enhanced when external sites are occupied by Mg(II).[40] This supports our findings that these sites are better stabilized by Mg(II)[37] rather than Ca(II).

Therefore, the difference of the two geometries (1J34 and 2A2Q) resides in the fact that, in 2A2Q, the central site is occupied by a Mg(II) which is only able to complete its coordination shell with water, for which the residence time has been measured to be on the order of 1 *μ*s.[8] This explains the greater distance between the two α-helices and the *ω*-loop (2A2Q case), namely, the interposition of a water layer. Therefore, one of the roles of the cations is to fix water, which in turn binds to the *ω*-loop through a network of H-bonds. By contrast, in FIXa 1J34, where all the central cation binding sites are occupied by Ca(II), interactions take place directly between metal cations and the lone pairs borne by its ligating oxygens. In this connection, it was recently demonstrated that water layers present either between separate domains of a protein or in between different proteins are in dynamic short time exchanges with the solvation water present in the environment.[50,51] Therefore, it is possible to make the hypothesis that F VIIa crystallized by Bajaj et al. (2A2Q), with interactions through water (i.e., water mediated interactions), could be present, free in the blood plasma since the latter is a highly hydrophilic environment. On the other hand, the GLA domain of F IXa crystallized by Shikamoto et al. (1J34) is found bound to a ligand (in this case, a snake venom protein); this binding could impose the “direct interactions with cations” geometry. It is, then, perhaps reasonable to suggest that this is the geometry that inserts deeply into the cell membrane, the interior of which is a highly hydrophobic environment. Indeed, the insertion within the membrane may expel water molecules from the inserted part of the GLA domain, imposing a switch between “physics at play”: going from a “through water structure” to a “direct electronic interactions structure”.

## Conclusion

In this contribution, we have used several theoretical tools to illustrate the relationship between the electronic structure of selected metal−cation complexes and the so-called hard and soft chemical behavior. The ELF analysis allows ranking cations according to their topological signature, namely, their ability to split their valence into subdomains (subvalent basins). Hard cations will not exhibit such a capability as soft cations do. The covalent character of the ligand−metal cation interaction is associated with a basin between the metal and the neighboring heavy atoms. That way, such a covalent interaction can be directly affected by the cations’ environment, and as we have seen, an electrostatic interaction, as in Mg(II) monodentate-like complexes, can become covalent, exhibiting an extra subvalent localization basin under specific stress conditions. This topological metal cation ranking has been shown to be relevant when compared to RVS and CSOV energy analyses, as well as in good qualitative agreement with local Fukui functions extracted from both QTAIM and ELF analyses.

Our integrated methodological approach uncovered a clear relationship between the underlying metal electron structure and the biological activity of enzymes such as vitamin-K-dependent coagulation factors. The present approach could then be extended to other biological systems involving metal cations. It is also important to point out that these results on metal cation−ligand complexes, such as the dissymmetry between internal and external carboxylate oxygen lone pairs, could also provide useful information for the design of new force fields.[7a,52] Moreover, applications to the problem of blood coagulation allowed us to address the Ca(II) vs Mg(II) selectivity in their interaction with GLA domains. Only the two X-ray structures, crystallized using the same Mg(II)/Ca(II) mix found under physiological conditions (1J34 and 2A2Q), have been investigated.

In the first one (1J34), Ca(II) has been shown to be more covalently bonded to ligand than Mg(II), enabling the creation of a “direct charge transfer network” between its subvalence and the carboxylate oxygen lone pairs. We showed that this concept could also uncover the distinct role of the two cation binding sites present in GLA domains. Being at the origin of the observed charge transfer network, calcium cations are mandatory in the central region to conserve the folding, whereas external binding sites are better stabilized by Mg(II), in agreement with the experiment, thanks to strong electrostatic interactions with the environment.

In the second structure (2A2Q), in which some interactions between metal cations and Gla are established through the presence of crystallized water molecules, we have shown that magnesium is able to form “an indirect charge transfer network” with malonates through its interaction with two highly polarized water molecules that are responsible for a structure of the $\omega$-loop similar to that observed in the first structure (1J34). To conclude, as the two crystal structures have been solved, they could be the two interconverting forms of the same system, each of them having a different water requirement. In both cases, the $\omega$-loop is present. This is required for the active enzyme, but the underlying physics is not the same. In the first structure, the folding is mainly due to electronic effects: Ca(II) is preferred in the five central binding sites. By contrast, the presence of water molecules is able to reverse the direct electronic selectivity of cations (Mg(II) could be present if interacting with two water molecules). This clearly indicates that dynamical solvent effects could play a key role in the observed structure of the domain. The connection between the two structures might be established through a series of molecular dynamics simulations. This work also emphasizes again[53] the importance of a “discrete” water molecule to understand the stability of biological systems, as the presence of a limited number of structured water molecules could be critical to obtain meaningful theoretical models.

**Acknowledgment.** The computations have been performed at the national IDRIS (F. 91403 Orsay, France), CRIHAN (F.76800 Saint-Etienne-du-Rouvray, France), and CINES (F. 34097 Montpellier, France) supercomputing centers. Some ELF computations have been run at the local CCRE centre at Université Pierre et Marie Curie, Univ Paris 6 (F. 75252 Paris CEDEX 05, France). Support from the French National Research Agency (ANR) on project LASIHMODo (BLAN08-3_312754) is acknowledged. This research was supported in part by the Intramural Research program of the NIH and NIEHS.

**Supporting Information Available:** (S1) Performance of the B3LYP functional. (S2) Effects of PCM solvation over the topology of two bidentate formate-cation complexes. (S3) CSOV computations (DFT level). This information is available free of charge via the Internet at http://pubs.acs.org/.

## References

(1) Furie, B.; Furie, B. C. *Cell* **1988**, *53*, 505–518.

(2) Prendergast, F. G.; Mann, K. G. *J. Biol. Chem.* **1977**, *252*, 840–850.

(3) van den Besselaar, A. M. H. P. *Blood Coagulation Fibrinolysis* **2002**, *13*, 19–23.

(4) Wang, S. X.; Hur, E.; Sousa, C. A.; Brinen, L.; Slivka, E. J.; Fletterick, R. J. *Biochemistry* **2003**, *42*, 7959–7966.

(5) Persson, E.; Ostergaard, A. *J. Tromb. Haemost.* **2007**, *5*, 1977–1978.

(6) (a) Lightstone, F. C.; Schwegler, E.; Hood, R. Q.; Gygi, F.; Galli, G. *Chem. Phys. Lett.* **2001**, *343*, 549–555. (b) Naor, M. M.; Van Nostrand, K.; Dellago, C. *Chem. Phys. Lett.* **2003**, *369*, 159–164. (c) Bako, I.; Hutter, J.; Palinkas, G. *J. Chem. Phys.* **2002**, *117*, 9838–9843. (d) Lightstone, F. C.; Schwegler, E.; Allesch, M.; Gygi, F.; Galli, G. *Chem. Phys. Chem.* **2005**, *6*, 1745–1749.

(7) (a) Piquemal, J.-P.; Perera, L.; Cisneros, G. A.; Ren, P.; Pedersen, L. G.; Darden, T. A. *J. Chem. Phys.* **2006**, *125*, 054511−1-7. (b) Babu, C. S.; Lim, C. *J. Phys. Chem. A* **2006**, *110*, 691–699.

(8) Neely, J.; Connick, R. *J. Am. Chem. Soc.* **1970**, *92*, 3476–3478.

(9) (a) Bader, R. F. W. *Atoms In Molecules: A Quantum Theory*; Oxford University Press: Oxford, U. K., 1990. (b) Matta, C. F.; Boyd, R. J. *The Quantum Theory of Atoms in Molecules: From Solid State to DNA and Drug Design*; Wiley-VCH: Weinheim, Germany, 2007. (c) Popelier, P. L. A. *Atoms In Molecules: An Introduction*; Prentice-Hall: Harlow, U. K., 2000.

(10) (a) Becke, A. D.; Edgecombe, K. E. *J. Chem. Phys.* **1990**, *92*, 5397–5403. (b) Silvi, B.; Savin, A. *Nature (London)* **1994**, *371*, 683–686.

(11) (a) Stevens, W. J.; Fink, W. *Chem. Phys. Lett.* **1987**, *139*, 15–22. (b) Bagus, P. S.; Illas, F. *J. Chem. Phys.* **1992**, *96*, 8962–8970.

(12) Silvi, B.; Fourré, I.; Alikani, M. E. *Chemie* **2005**, *136*, 855–879.

(13) Silvi, B. *J. Phys. Chem. A* **2003**, *107*, 3081–3085.

(14) Martín Pendás, A.; Francisco, E.; Blanco, M. A. *Chem. Phys. Lett.* **2008**, *454*, 396–403.

(15) (a) Gillespie, R. J.; Popelier, P. L. A. *Chemical Bonding and Molecular Geometry*; Oxford University Press: Oxford U. K., 2001. (b) Gillespie, R. J.; Robinson, E. A. *J. Comput. Chem.* **2007**, *34*, 396–407.

(16) Gillespie, R. J.; Noury, S.; Pilme, J.; Silvi, B. *Inorg. Chem.* **2004**, *43*, 3248–3256.

(17) Piquemal, J.-P.; Pilmé, J.; Parisel, O.; Gérard, H.; Fourré, I.; Bergès, J.; Gourlaouen, C.; de la Lande, A.; van Severen, M. C.; Silvi, B. *Int. J. Quantum Chem.* **2008**, *108*, 1951–1969.

(18) (a) Pilmé, J.; Piquemal, J.-P. *J. Comput. Chem.* **2008**, *29*, 1440–1449. (b) Fukui, K.; Yonezawa, T.; Shingu, H. *J. Chem. Phys.* **1952**, *20*, 722–725. Fukui, K.; Yonezawa, T.; Nagata, C.; Shingu, H. *J. Chem. Phys.* **1954**, *22*, 1433–1442. (c) Woodward, R. B.; Hoffmann, R. *The Conservation of Orbital Symmetry*; Chemie: Weinheim, Germany, 1970. (d) Parr, R. G.; Yang, W. *Density Functional Theory of Atoms and Molecules*; Oxford University Press: New York, 1989. (e) Morell, C.; Grand, A.; Toro-Labbé, A. *J. Phys. Chem. A* **2005**, *109*, 205–212. (f) Morell, C.; Grand, A.; Toro-Labbé, A. *Chem. Phys. Lett.* **2006**, *425*, 342–346. (g) Cioslowski, J.; Martinov, M.; Mixon, S. T. *J. Phys. Chem.* **1993**, *97*, 10948–10951. (h) Contreras, R. R.; Fuentealba, P.; Galván, M.; Pérez, P. *Chem. Phys. Lett.* **1999**, *304*, 405–413. (i) Bulat, F. A.; Chamorro, E.; Fuentealba, P.; Toro-Labbé, A. *J. Phys. Chem. A* **2004**, *108*, 342–349. (j) Tiznado, W.; Chamorro, E.; Contreras, R.; Fuentealba, P. *J. Phys. Chem. A* **2005**, *109*, 3220–3224.

(19) (a) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789. (b) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652. (c) Burda, J. V.; Šponer, J.; Hobza, P. J. *J. Phys. Chem.* **1996**, *100*, 7250–7256. (d) Russo, N.; Toscano, M.; Grand, A. *J. Phys. Chem. A* **2003**, *107*, 11533–11538.

(20) *Jaguar 6.5*; Schrodinger Inc.: Portland, OR, 2005.

(21) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 299–310.

(22) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650–654.

(23) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, Revision C.02; Gaussian Inc.: Wallingford, CT, 2007.

(24) Noury, S.; Krokidis, X.; Fuster, F.; Silvi, B. *J. Comput. Chem.* **1999**, *23*, 597–604. The modified TopMod90 program (named TopChem) is available upon request. See the following Web site for details http://www.lct.jussieu.fr/pagesperso/pilme (accessed 01/15/2010).

(25) Piquemal, J.-P.; Marquez, A.; Parisel, O.; Giessner-Prettre, C. *J. Comput. Chem.* **2005**, *26*, 1052–1062.

(26) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery Jr, J. A. *J. Comput. Chem.* **1993**, *14*, 1347–1363.

(27) Ratcliffe, J. V.; Furie, B.; Furie, B. C. *J. Biol. Chem.* **1993**, *268*, 24339–24345.

(28) de Courcy, B.; Gresh, N.; Piquemal, J.-P. *Interdiscip. Sci. Comput. Life Sci.* **2009**, *1*, 55–60.

(29) Parr, R. G.; Pearson, R. G. *J. Am. Chem. Soc.* **1983**, *105*, 7512–7516.

(30) Xu, X.; Zhang, L.; Shen, D.; Wu, H.; Peng, L.; Li, J. *J. Biol. Inorg. Chem.* **2009**, *14*, 559–571.

(31) Perera, L.; Foley, C.; Darden, T. A.; Stafford, D.; Mather, T.; Esmon, C. T.; Pedersen, L. G. *Biophys. J.* **2000**, *79*, 2925–2643.

(32) Hoffman, M. *J. Thromb. Thrombolysis* **2003**, *16*, 17–20.

(33) Roberts, H. R.; Hoffman, M.; Monroe, D. M. *Semin. Thromb. Hemost.* **2006**, *32* (Suppl. 1), 32–38.

(34) McDonald, J. F.; Shah, A. M.; Schwalbe, R. A.; Kisiel, W.; Dahlbäck, B.; Nelsestuen, G. L. *Biochemistry* **1997**, *36*, 5120–5127.

(35) Soriano-Garcia, M.; Padmanabhan, K.; de Vos, A. M.; Tulinsky, A. *Biochemistry* **1992**, *31*, 2554–2566.

(36) Gopinath, S. C. B.; Shikamoto, Y.; Mizuno, H.; Kumar, P. K. R. *Biochem. J.* **2007**, *405*, 351–357.

(37) Sekiya, F.; Yamashita, T.; Atoda, H.; Komiyama, Y.; Morita, T. *J. Biol. Chem.* **1995**, *270*, 14325–14331.

(38) Banner, D. W.; D'Arcy, A.; Chene, C.; Winkler, F. K.; Guha, A.; Konigsberg, W. H.; Nemerson, Y.; Kirchhofer, D. *Nature (London)* **1996**, *380*, 41–46.

(39) Shikamoto, Y.; Morita, T.; Fujimoto, Z.; Mizuno, H. *J. Biol. Chem.* **2003**, *278*, 24090–24094.

(40) Mizuno, H.; Fujimoto, Z.; Atoda, H.; Morita, T. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 7230–7234.

(41) Bajaj, S. P.; Schmidt, A. E.; Agah, S.; Bajaj, M. S.; Padmanabhan, K. *J. Biol. Chem.* **2006**, *281*, 24873–24888.

(42) Falls, L. A.; Furie, B. C.; Jacobs, M.; Furie, B.; Rigby, A. C. *J. Biol. Chem.* **2001**, *276*, 23895–23902.

(43) Ohbuko, Y. Z.; Tajkhorshid, E. *Structure* **2008**, *16*, 72–81.

(44) Sunnerhagen, M.; Forsén, S.; Hoffrén, A.-M.; Drakenberg, T.; Teleman, O.; Stenflo, J. *Nat. Struct. Biol.* **1995**, *2*, 504–509.

(45) Taboureau, O.; Olsen, O. H. *Eur. Biophys. J.* **2007**, *36*, 133–144.

(46) Davis, C. H.; Deerfield, D., II; Stafford, D. W.; Pedersen, L. G. *J. Phys. Chem. A* **2007**, *111*, 7257–7261.

(47) Huang, M.; Furie, B. C.; Furie, B. *J. Biol. Chem.* **2004**, *279*, 14338–14346.

(48) Huang, M.; Rigby, A. C.; Morelli, X.; Grant, M. A.; Huang, G.; Furie, B.; Seaton, B.; Furie, B. C. *Nat. Struct. Biol.* **2003**, *10*, 751–756.

(49) Sekiya, F.; Yoshida, M.; Yamashita, T.; Morita, T. *J. Biol. Chem.* **1996**, *271*, 8541–8544.

(50) Lin, J.; Balabin, I. A.; Beratan, D. N. *Science* **2005**, *310*, 1311–1313.

(51) de La Lande, A.; Marti, S.; Parisel, O.; Moliner, V. *J. Am. Chem. Soc.* **2007**, *129*, 11700–11707.

(52) (a) Gresh, N.; Cisneros, G. A.; Darden, T. A.; Piquemal, J.-P. *J. Chem. Theory Comput.* **2007**, *3*, 1960–1986. (b) Piquemal, J.-P.; Chevreau, H.; Gresh, N. *J. Chem. Theory Comput.* **2007**, *3*, 824–837. (c) Piquemal, J.-P.; Cisneros, G. A.; Reinhardt, P.; Gresh, N.; Darden, T. A. *J. Chem. Phys.* **2006**, *124*, 104101−1-12.

(53) de Courcy, B.; Piquemal, J.-P.; Garbay, C.; Gresh, N. *J. Am. Chem. Soc.* **2010**, *132*, 3312−3320.

CT100089S

JCTC Journal of Chemical Theory and Computation

# Bonding in Classical and Nonclassical Transition Metal Carbonyls: The Interacting Quantum Atoms Perspective

Davide Tiana,[†,⊥] E. Francisco,[‡] M. A. Blanco,[‡] P. Macchi,[§] Angelo Sironi,[†] and A. Martín Pendás*[,‡]

*Department of Structural Chemistry and Inorganic Stereochemistry, University of Milan, Via Venezian 21, 20133 Milan, Italy, Departamento de Química Física y Analítica, Facultad de Química, Universidad de Oviedo, 33006-Oviedo, Spain, and Department of Chemistry and Biochemistry, University of Bern, Switzerland*

Received December 9, 2009

**Abstract:** Chemical bonding in simple transition metal carbonyls is examined under the interacting quantum atoms approach (IQA), which provides an energetic viewpoint within the quantum theory of atoms in molecules (QTAIM). We have studied both classical and nonclassical isoelectronic series of complexes, with different coordinations and geometries and studied the evolution of the IQA interatomic interactions, using several levels of theory. Our results in classical carbonyls are compatible with the standard Dewar−Chatt−Duncanson model, although multicenter bonding may have an important role in some complexes. The increase (decrease) in the CO distance upon bonding is faithfully coupled to a decrease (increase) in the CO covalent energy, although the main energetic change in the CO moiety is electrostatic and due to charge transfer and/or polarization of its electron density. The metal−ligand interaction energy is dominated by covalent effects and depends strongly on the total net charge of the complex, being larger for negatively charged molecules, where $\pi$-back-donation is very important. The electrostatic (ionic-like) metal−ligand interaction energy is small in general, although it becomes more and more stabilizing with increasing coordination number.

## 1. Introduction

The metal (M) carbonyl (CO) interaction has probably sparked more interest than any other in metallorganic chemistry. This importance stems from its paradigmatic role in chemical bonding theory as well as in surface chemistry and catalysis.

As bonding is regarded, the M−CO interaction is generally gauged against the classical model proposed by Dewar, Chatt, and Duncanson (DCD) in 1951.[1,2] Succintly, a synergistic interaction is proposed to occur between $\sigma$ charge donation from the CO highest occupied molecular orbital (HOMO) and a consequent $\pi$-back-donation from the M d orbitals to the CO lowest unoccupied molecular orbital (LUMO). Since its proposal, most theoretical works, using very many interpretation tools that range from Kitaura−Morokuma energetic decompositions to Natural Bond Orbital (NBO) analyses or real space techniques like those based on the Quantum Theory of Atoms in Molecules (QTAIM) or the use of the Electron Localization Function (ELF), have corroborated the essence of the DCD model.[3−8]

It is generally thought, for instance, that although the $\sigma$ donation is larger than the $\pi$-back-donation, their energetic bonding role is reversed,[9] and the back-donation contribution is actually dominant.[10,11] From a simple molecular orbital (MO) perspective, the CO HOMO is its $5\sigma$ orbital, while the LUMO is a relatively low lying $2\pi^*$ function. Similarly, the metal frontier orbitals are the (crystal field splitted) d orbitals. The success of the DCD model is based on its simple, straightforward qualitative predictions. For instance,

* Corresponding author e-mail: angel@fluor.quimica.uniovi.es.

[†] University of Milan.

[‡] Universidad de Oviedo.

[§] University of Bern.

[⊥] Previous address: Departamento de Química Física y Analítica, Facultad de Química, Universidad de Oviedo, 33006-Oviedo, Spain.

10.1021/ct9006629 © 2010 American Chemical Society
Published on Web 03/19/2010

it has been repeatedly reported that, for the CO bond, the $5\sigma$ orbital is either nonbonding,[12] or slighty antibonding,[13] while the nature of the $2\pi^*$ is definitely antibonding. In this way, the flow of electrons into the latter easily rationalizes the weakening of the CO bond, with its consequent lengthening[14] and reduction of stretching frequencies.[15] These two effects are standardly used to quantify the back-donation and evaluate the (reduced) CO bond strength.

A number of other experimental facts stem from similar DCD MO arguments. Back-bonding, for instance, clearly depends on the d$-2\pi^*$ energy gap,[16] which in turn is related to the electron−nuclear attraction and then, for an isoelectronic series, to the M atomic number. Since $\pi$-back-bonding implies a withdrawal of electronic density from M to the CO ligand, it is related to the M ionization potential. Further, $\pi$-back-donation being a shorter ranged interaction than $\sigma$ donation, it turns out to be relatively sensible to the M covalent radius. Thus, relativistic effects and lanthanide contractions play a non-negligible role in its magnitude.[17]

The model illustrated so far was first questioned in the 1970s when, for the first time, the metal carbonyl cations $\mathrm{Cu(CO)}_n^+$, $n = 1$ and 2, were synthesized.[18] These examples were followed by other homoleptic noble metal carbonyls like $\mathrm{Ag(CO)}_n^+$, $n = 1$ and 2, or $\mathrm{Au(CO)}_2^+$. Their common feature was a higher CO stretching frequency than that found in free CO (2143 $\mathrm{cm}^{-1}$), so they were called "nonclassical" by Strauss et al.[17] The effect was ascribed to the absence of $\pi$-back-donation, such that the remaining weak $\sigma$ donation removed density from the antibonding $5\sigma$ orbital, this resulting in a shorter CO bond with a larger force constant. Criticisms about the arbitrariness of such an explanation soon arrived,[19,20] and for instance, the antibonding character of the $5\sigma$ MO was put into question. Not only did a few studies demonstrate that this orbital was not antibonding, but also that the CO bond stiffenning was correctly reproduced by just modeling the electric field induced by the M cation. Slowly, an image in which the density redistribution induced by this field increased the covalency of the CO bond emerged.[21,22]

It is now relatively clear that it is not so easy to correlate CO stretching frequencies or force constants to M−CO back-donation due to mode coupling, that Strauss' inverse correlation between M−C and C−O distances may fail,[21] and that back-bonding is basically dependent on the M−CO distance, so that a particular onset distance exists for nearly every system that may be larger or smaller than its particular equilibrium geometry.[23] It has even been shown that the amount of donation−back-donation cannot be used as an indicator of binding energies.[24]

Real space analyses of chemical bonding in transition metal carbonyls have also been commonplace, in both the QTAIM and ELF flavors.[20,23,7,5,6] The MC bond is usually characterized by large positive laplacians at the bond critical point (bcp), with relatively large delocalization indices ($\delta^{AB}$) for the MC pair. According to the QTAIM indicators, the MC bond has characteristics similar to those found in dative interactions of main group elements. Attempts to validate the DCD model have also been made either by partitioning densities into $\sigma$ and $\pi$ contributions,[7] by showing how the $\delta^{\mathrm{MO}}$ correlates, as expected, with back-bonding,[5,25] or by using the domain averaged Fermi hole (DAFH) technique,[26] as introduced by Ponec.[27,28] The DCD model has also been examined in terms of real space valence charge concentrations for the metal−olefin link.[29] The solid theoretical foundation of these real space techniques is making them increasingly popular in the field of chemical bonding in transition metal (TM) chemistry.[8] However, a real space energetic image of these important bonds is lacking, and our interacting quantum atoms approach (IQA) may clearly fill this gap.

IQA[30−34] provides a theory of cohesion within the QTAIM by extending the domain partitioning of one-electron observables to interelectron repulsions. By doing so, we get an exact, chemically appealing partition of the molecular energy that may be recast into an easy to comprehend language. Within IQA, binding is the result of a competition between atomic deformation, which is an analogue of the classical promotion energy needed for an atom to get bonded to another, and pairwise additive interatomic interaction energies. The latter are made up of classical (or ionic-like) and exchange-correlation, purely quantum mechanical (or covalent-like) components. IQA has now been applied to provide a real space energetic view of a wide number of problems,[35−37] to shed some light into core concepts of the QTAIM,[38] and to propose statistical images of the chemical bond.[39−42] Recently, we have shown how to generalize its framework to wave functions containing effective core potentials (ECPs),[43] opening a window to examine chemical bonding issues in TM chemistry.

In this paper, we will examine the energetics of simple metal carbonyls under the IQA light, paying particular attention to the possible difference between classical and nonclassical systems. We will also show how many of the accepted energy features of the DCD model are translated into state of the art, orbital invariant real space analyses.

The layout of the paper is as follows. First, a brief summary of the IQA procedure will be presented, followed by a description of the computational details of our calculations. Then, we will describe our results in a number of classical and nonclassical carbonyls. We will end with some conclusions.

## 2. Brief IQA Survey and Summary

Let us succintly introduce a survey of the IQA parlance. Full accounts may be found elsewhere.[30−34] Let us start with a QTAIM partition of the space into atomic domains.[44] Then, at any molecular geometry,

$$\begin{aligned} E = & \sum_A (T^{\mathrm{A}} + V_{\mathrm{en}}^{AA} + V_{\mathrm{ee}}^{AA}) \\ & + \sum_{A>B} (V_{\mathrm{nn}}^{AB} + V_{\mathrm{en}}^{AB} + V_{\mathrm{ne}}^{AB} + V_{\mathrm{ee}}^{AB}) \\ = & \sum_A E_{\mathrm{self}}^{A} + \sum_{A>B} E_{\mathrm{int}}^{AB} \end{aligned} \tag{1}$$

where $A \equiv \Omega_A$ is the atomic basin of nucleus $A$; $T^{\mathrm{A}}$ is its atomic kinetic energy; and $V_{\mathrm{en}}$, $V_{\mathrm{ne}}$, $V_{\mathrm{ee}}$, and $V_{\mathrm{nn}}$ are the potential energies describing the several pair interactions between the electrons and nuclei that reside in basins $A$ and $B$, in an easy to decode terminology. Now, IQA uses ideas borrowed from McWeeny's[45] electronic separability to gather the above terms using chemical insight. All intrabasin terms are added to define the self-energy of a quantum atom (or group) $E^{A}_{\mathrm{self}}$, while the interbasin ones are gathered in the interaction energy between pairs of atoms, $E^{AB}_{\mathrm{int}}$. This interaction energy may be further partitioned into a classical component, $V^{AB}_{\mathrm{cl}}$, obtained by adding $V_{\mathrm{en}}$, $V_{\mathrm{ne}}$, and $V_{\mathrm{nn}}$; the Coulombic part of $V_{\mathrm{ee}}$, $V^{AB}_{\mathrm{C}}$; and a quantum mechanical, nonclassical, or exchange correlation term $V^{AB}_{\mathrm{xc}} = V^{AB}_{\mathrm{ee}} - V^{AB}_{\mathrm{C}}$. This may always be done, so $E^{AB}_{\mathrm{int}} = V^{AB}_{\mathrm{cl}} + V^{AB}_{\mathrm{xc}}$. When particular energetic references exist for the quantum groups, $E^{A,0}_{\mathrm{self}}$, we define atomic deformation energies, $E^{A}_{\mathrm{def}} = E^{A}_{\mathrm{self}} - E^{A,0}_{\mathrm{self}}$, such that molecular binding is the result of a competition between terms with the same order of magnitude: group deformation (which is usually positive) and intergroup interaction (overall negative):

$$E_{\mathrm{bind}} = \sum_{A} E^{A}_{\mathrm{def}} + \sum_{A>B} (V^{AB}_{\mathrm{cl}} + V^{AB}_{\mathrm{xc}}) \qquad (2)$$

We have shown[32,34] that the classical and exchange-correlation interaction components are associated with the conventional notions of ionicity and covalency, so IQA translates quantum-mechanical energetic quantities into standard chemical concepts.

The formalism is immediately extended if several quantum groups are gathered to form functional groups $\mathscr{G}$, $\mathscr{H}$, etc.[32,34] Equation 1 applies for groups, too, if $E^{\mathscr{G}}_{\mathrm{self}}$ is defined to contain all intragroup energetic components, and also eq 2 if whole-group references are used to define $E^{\mathscr{G}}_{\mathrm{def}}$. In this work, it will be useful to consider each carbonyl ligand (L) as a quantum group, such that $\Omega_{\mathrm{L}} = \Omega_{\mathrm{C}} \cup \Omega_{\mathrm{O}}$. In this way, we may discuss either the ML link interaction as a whole or each of its MC and MO components. Also, it is advantageous to discuss the changes in the ligand-related properties with respect to those in the isolated CO molecule, introducing the notation $\Delta X = X$ (coordinated L) $- X$ (free L). Using this notation, with free CO as a reference, the ligand deformation energy can be written as $E^{\mathrm{L}}_{\mathrm{def}} = \Delta E^{\mathrm{C}}_{\mathrm{self}} + \Delta E^{\mathrm{O}}_{\mathrm{self}} + \Delta E^{\mathrm{CO}}_{\mathrm{int}}$. Binding of the $n$ CO ligands to the metal will thus comprise $n$ ML interactions, $E^{\mathrm{ML}}_{\mathrm{int}} = E^{\mathrm{MC}}_{\mathrm{int}} + E^{\mathrm{MO}}_{\mathrm{int}}$, $n$ of the above-mentioned ligand deformation energies measuring the cost in coming from the free CO state into the complex, plus the much smaller LL interactions, and the metal deformation energy $E^{\mathrm{M}}_{\mathrm{def}}$.

## 3. Computational Details

IQA partitioning, necessarily based on numerical integration techniques, is computationally intensive[30] and requires well-defined first and second order density matrices. Thus, only Hartree−Fock (HF) or variational multiconfigurational techniques may be used. Since the literature on M−CO systems is huge, much is known about the performance and limitations of methods with different levels of approximation for the treatment of electron correlation. For instance, Sherwood and Hall[46] showed that 97% the total energy is recovered at the HF level for $Cr(CO)_6$. At this moment, it is well-known that MP2 results paralell those trends found at the SCF level, with only quantitative differences. Single determinant approaches give rise to lower binding energies, thus longer MC and shorter CO distances,[47] and to underestimated $\pi$-back-donating but reasonable $\sigma$-donating effects.[24]

Less is known about the performance of DFT techniques as chemical bonding is regarded in these compounds. The B3LYP functional seems to overestimate back-donation,[8] which has turned out to be quite sensitive to the exchange-correlation functional.[21] Similarly, both BLYP and B3LYP overestimate binding energies giving rise to long MC distances in $CuCO^{+}$ and $CuCO^{2+}$.[48] All in all, whenever a multiconfigurational calculation cannot be undertaken due to computational constraints, a simple HF wave function provides a reasonable account of the major binding forces acting on simple transition metal carbonyls. Thus, we present here a combination of HF, CASSCF, and DFT (approximate) results.

All electronic structure calculations were performed with GAMESS.[49] H and main group elements were modeled with a 6-31G(d,p) basis set, while standard Hay and Wadt small core relativistic ECPs together with their standard basis sets (3s4s1s3p1p1p4d1d) were used for the transition metals.[50] IQA analyses were done with our PROMOLDEN code, and ECPs were treated according to our previously published protocol,[43] with M core densities obtained from 3-21G basis sets added in the computation of appropriate interatomic surfaces. It is important to recall that deformation energies are not accessible with our protocol for those quantum atoms bearing ECPs.[43] In the present case, we thus have access to $E^{\mathrm{L}}_{\mathrm{def}}$, but not to $E^{\mathrm{M}}_{\mathrm{def}}$, so we will only briefly consider the ligand deformation energies in a subset of our calculations, focusing the discussion on the ML interactions.

We have studied several ideal low-spin isoelectronic series at the HF level: The $d^{10}$ $T_d$ $[Fe(CO)_4]^{2-}$, $[Co(CO)_4]^{-}$, $Ni(CO)_4$, $[Cu(CO)_4]^{+}$, and $Pd(CO)_4$ systems, together with their $d^8$ $D_{4h}$ $[Ni(CO)_4]^{2+}$ and $[Pd(CO)_4]^{2+}$ planar counterparts; the $d^8$ $D_{3h}$ $[Mn(CO)_5]^{-}$, $Fe(CO)_5$, $[Co(CO)_5]^{+}$, and $Ru(CO)_5$ pentacarbonyls; and the $d^6$ $O_h$ $[Ti(CO)_6]^{2-}$, $[V(CO)_6]^{-}$, $Cr(CO)_6$, $[Mn(CO)_6]^{+}$, and $[Fe(CO)_6]^{2+}$ species. Homoleptic noble metal cations $[M(CO)_n]^{+}$, M = (Cu, Ag, Au) with $n$ = 1 and 2, have also been studied at the CASSCF//MP2 level with an active space comprising the five valence M d orbitals supplemented with the five outer orbitals of each carbonyl and four suitable low-lying virtuals. Due to the lack of appropriate basis sets, no core density was added in the CASSCF calculations. In order to test how IQA performs when using a Kohn−Sham determinant to approximately construct first- and second-order density matrices, we also performed some test DFT calculations at MP2 geometries for $[Ag(CO)_2]^{+}$ and $[Au(CO)_2]^{+}$. Several functionals have been used: BLYP; BLYP-LC; and the M06-l, M06, and M06-HF series, characterized by 0%, 27%, and 100% HF exchange, respectively. The $d^{10}$ $T_d$ molecules were also computed at the M06 level. To ascertain the role of electrostatic effects in nonclassical carbonyls, we have also computed the $HCO^{+}$ and $COH^{+}$ systems at the equivalent CASSCF level. Wave functions for the isolated CO ligand

***Table 1.*** Basic Geometric and QTAIM Integrated Properties Together with IQA Interactions for the $[Fe(CO)_4]^{2-}$, $[Co(CO)_4]^{-}$, $Ni(CO)_4$, $[Cu(CO)_4]^{+}$, and $Pd(CO)_4$ $d^{10}$ $T_d$ Tetracarbonyls[a]

| M | Fe | Co | Ni | Cu | Pd |
|---|---|---|---|---|---|
| $d$(MC) | 1.735 | 1.766 | 1.924 | 2.296 | 2.229 |
| $\Delta d$(CO) | 0.049 | 0.024 | 0.002 | −0.010 | 0.000 |
| $\Delta\nu$ | −491 | −295 | −64 | 90 | −32 |
| $Q^{M}$ | 0.282 | 0.189 | 0.122 | 0.802 | 0.064 |
| $Q^{L}$ | −0.570 | −0.297 | −0.031 | 0.050 | −0.016 |
| $\Delta Q^{C}$ | −0.439 | −0.234 | −0.086 | −0.035 | −0.025 |
| $\delta^{MC}$ | 1.347 | 1.153 | 0.798 | 0.313 | 0.623 |
| $\delta^{MO}$ | 0.198 | 0.166 | 0.097 | 0.024 | 0.075 |
| $\delta^{CC}$ | 0.130 | 0.084 | 0.039 | 0.012 | 0.016 |
| $\Delta\delta^{CO}$ | −0.318 | −0.231 | −0.046 | 0.037 | −0.056 |
| $E_{int}^{ML}$ | −0.343 | −0.277 | −0.192 | −0.070 | −0.131 |
| $E_{int}^{MC}$ | −0.215 | −0.187 | −0.142 | 0.122 | −0.098 |
| $\Delta E_{int}^{CO}$ | 0.374 | 0.220 | 0.212 | 0.050 | 0.044 |
| $V_{cl}^{ML}$ | −0.008 | 0.011 | 0.005 | −0.005 | 0.002 |
| $V_{cl}^{MC}$ | 0.100 | 0.084 | 0.045 | 0.185 | 0.029 |
| $\Delta V_{cl}^{CO}$ | 0.312 | 0.180 | 0.224 | 0.068 | 0.040 |
| $V_{xc}^{ML}$ | −0.334 | −0.288 | −0.197 | −0.067 | −0.134 |
| $V_{xc}^{MC}$ | −0.314 | −0.271 | −0.187 | −0.065 | −0.127 |
| $\Delta V_{xc}^{CO}$ | 0.063 | 0.041 | −0.011 | −0.017 | 0.004 |
| $E_{def}^{L}$ | 0.236 | 0.179 | 0.110 | 0.064 | 0.087 |

[a] HF data in atomic units, except distances in Å and frequencies in $cm^{-1}$. Parameters for the isolated CO molecule are as follows: $d(CO) = 1.114$, $\nu(CO) = 2439$, $Q^{C} = 1.403$, $\delta^{CO} = 1.508$, $E_{int}^{CO} = -2.120$, $V_{cl}^{CO} = -1.706$, and $V_{xc}^{CO} = -0.415$. Recall that $\Delta X = X$ (coordinated L) $-$ $X$ (free L).

have been obtained for comparison purposes at all the computational levels previously described.

PROMOLDEN integrations have been performed using typically tight parameters, truncated at $l_{max} = 10$, with 631-point radial and 5810-point Lebedev angular grids. $\beta$-spheres up to $l = 6$ with radii equal to 90% of the distance from the nuclear position to the closest bcp have been used, and 431 radial and 534 Lebedev angular grid points have been selected for them. These are fairly standard computational conditions for IQA calculations[32] that usally provide interactions converged to about 1 kcal/mol.

Since IQA interaction energies vary in wide ranges, we will use atomic units for them in all of our tables. However, some of the arguments in the text will refer to relevant differences among them. In those cases, we will shift to kilocalories per mole to provide more chemically meaningful quantities.

## 4. Bonding in the $T_d$ M(CO)$_4$ Systems

Let us examine the isoelectronic $d^{10}$ $T_d$ Fe, Co, Ni, Cu, and Pd tetracarbonyls from our IQA perspective. We will start with a brief analysis of some standard QTAIM integrated quantities. Table 1 contains topological charges, $Q$, and delocalization indices, $\delta$, for our systems, including the isolated CO ligand.

Although these kind of results are known, it should be noticed that all the metals bear small positive topological charges, meaning that, in the most simple DCD model, back-bonding should be extremely large in the Fe compound, for instance. The geometric correlations of the model, as briefly explained in the Introduction, do also come out easily from the computed data. For instance, the changes in the CO stretching frequencies upon bonding, the changes in $d$ (CO), and the total net charge of the ligand correlate among each other. Notice that the total charge of the ligand is mostly absorbed by the C atom, so charge transfer is fairly localized in the MC region.

Similar insight is obtained from electron delocalization properties, like the delocalization index, $\delta$, a measure of bond order in real space. For 3d metals, the MC bond order decreases as the MC distance increases, as expected. In line with DCD, this increase is coupled to a lengthening of the CO distance and a decrease in the CO bond order. Not so obviously, however, it is also clear that $\delta^{CO}$ does also correlate with the overall CO (L) polarization. The only system in which the L net charge is positive is $[Cu(CO)_4]^{+}$, meaning that charge transfer goes from the ligand to the metal in a $\sigma$-like fashion. This is also a possible nonclassical carbonyl, with smaller $d$(CO) and larger $\delta^{CO}$ than those found in free CO. Polarization of the CO ligand as induced by the metal is clearly related to the CO bond covalency, as pointed out by Lupinetti and co-workers.[23] Finally, another interesting point is related to the non-negligible $\delta^{CC}$ value that exists between adjacent L's in the Co and Fe compounds. This points toward an important multicenter character of the M−L bonding in those cases where back-bonding is also deemed important, i.e., in electron-rich compounds. As we will show with our DAFH analyses, this is true and should be understood as one limitation of the DCD view.

The energetic view provided by IQA enlightens the above comments. Let us start with a fine-grained view, by examining the MC and MO IQA quantities. $E_{int}^{MC}$ is large and splits the $T_d$ systems into two categories of negative and positive total MC interaction. As we will see, this is related to the total $Q^{L}$ charge, and positive MC total interactions will become common for other stoichiometries. The $V_{cl}^{MC}$ contribution to the MC interaction is destabilizing, due basically to the positive net charge at the metal site. However, its particular value is the result of a complex balance among the MC distance, the positive net M and C charges, and the polarization of the charge distribution. The covalent contribution to the MC bond, provided by $V_{xc}^{MC}$, follows the total net charge of the complex. Its value in the Fe, Co, and Ni compounds, about −200 kcal/mol in the first, is considerably large if we compare it to the −260 kcal/mol value obtained for the free CO ligand, $V_{xc}^{CO}$, a formal triple bond. MC covalency is the basic stabilizing interaction in this series, and even in the Cu case, where back-bonding is thought to have a minor role, it amounts to about −40 kcal/mol. The MO interactions may be obtained from the table by subtraction (MO = ML − MC) and deserve similar comments. $E_{int}^{MO}$ is obviously stabilizing, controlled by the negative electrostatic component which may be faithfully approximated by a point charge contribution, and its range of variation is smaller. Just as delocalization between the metal and the oxygen atom of each carbonyl provides a real space measure[25] of the relative intensity of $\pi$-back-donation, $V_{xc}^{MO}$ gives us its energetic signature. It is negligible in $[Cu(CO)_4]^{+}$, about 1 kcal/mol, and 10 times larger in the $[Fe(CO)_4]^{2-}$ anion. Notice that $V_{xc}^{MO}$ is always smaller than 7% $V_{xc}^{MC}$: this property should not be interpreted as a direct measure of the total energetics associated with back-donation, which does also include the C atom of the ligand, but rather as an isolated energetic signature, not affected by the covalent contribution of $\sigma$ donation.

Moving to group interactions, the total $E_{\mathrm{int}}^{\mathrm{ML}} = E_{\mathrm{int}}^{\mathrm{MC}} + E_{\mathrm{int}}^{\mathrm{MO}}$ is always negative, although in the Cu case it is relatively small, −44 kcal/mol. Notice that its classical component is small, its absolute value not exceeding 7 kcal/mol, and that it oscillates from positive to negative. Thus, even if each of the MC and MO electrostatic terms may be large, the polarization pattern of L conspires to overall small $V_{\mathrm{cl}}^{\mathrm{ML}}$ values such that in the end it is covalency, not electrostatics, that governs the ML bonding. Continuing with the group description of the CO ligands, their energetic change upon bonding deserves comment. The overall CO deformation energy, $E_{\mathrm{def}}^{\mathrm{L}}$, is positive (as found in most of the systems examined by IQA up to now) and is clearly correlated to the total net charge of the ligand, $Q^{\mathrm{L}}$. This is expected, since, in the case of highly heteropolar links, self-energies are controlled by ionization costs. Notice that the deformation energy in the Cu complex, with a rather low positive net charge, is relatively small, about 40 kcal/mol. In all the cases examined, deformation of the ligand is dominated by $\Delta E_{\mathrm{int}}^{\mathrm{CO}}$; we will concentrate on its components and will not consider $E_{\mathrm{def}}^{\mathrm{L}}$'s themselves again.

Regarding the different properties involved in the CO deformation upon coordination, $\Delta d^{\mathrm{CO}}$, $\Delta\delta^{\mathrm{CO}}$, and $\Delta V_{\mathrm{xc}}^{\mathrm{CO}}$ are quite linearly correlated, whereas $\Delta V_{\mathrm{cl}}^{\mathrm{CO}}$ is not. In agreement with previous IQA knowledge, and also with the results by Lupinetti and co-workers,[23] the CO distance responds basically to changes in covalency. These coordinated CO bonds display $V_{\mathrm{xc}}^{\mathrm{CO}}$ values smaller than those in free CO (except in the Cu complex), so ML bonding decreases the covalency of the CO bond. Although $V_{\mathrm{cl}}^{\mathrm{CO}}$ shows a relatively complex pattern, the simple $Q^{\mathrm{C}}Q^{\mathrm{O}}/d^{\mathrm{CO}}$ point charge term correlates rather well with it. In general, depolarization of a bond leads to a decrease in its electrostatic contribution, as may be rationalized from the smaller value of $(q - \varepsilon)(-q + \varepsilon)$ with respect to $-q^2$ in Coulomb's law when charge is transferred from one point charge to the other in an ionic pair.

We also note that, although in the Cu compound the covalent CO interaction energy is comparable but smaller than that in the free CO molecule, its electrostatic interaction is considerably (by about 60 kcal/mol) larger. The combination of both facts justifies its $d$ (CO), 0.01 Å smaller than the free value. We want to stress that $\Delta d$ (CO), $Q^{\mathrm{L}}$, $\Delta\delta^{\mathrm{CO}}$, $E_{\mathrm{int}}^{\mathrm{MC}}$, and $\Delta V_{\mathrm{xc}}^{\mathrm{CO}}$ all change sign in this complex with respect to the rest of the series, so the overall positive charge of the species has a big impact on its detailed energetics.

The behavior of the neutral Pd complex is also noteworthy, at least when compared to the also neutral Ni case. This is probably related to its large ionization potential that justifies its smaller $Q^{\mathrm{M}}$ and the relatively small change in the net charges that it induces on the CO ligand. However, the effect of its diffuse d shell on the ligand is not small, and the CO moiety suffers a rather big density polarization that increases its classical attraction.

***Table 2.*** DFT M06 Geometric and QTAIM Integrated Properties for the $T_d$ Complexes[a]

| M | Fe | Co | Ni | Cu | Pd |
|---|---|---|---|---|---|
| $d$(MC) | 1.745 | 1.764 | 1.843 | 2.038 | 2.081 |
| $\Delta d$(CO) | 0.055 | 0.029 | 0.006 | −0.010 | 0.004 |
| $Q^{\mathrm{M}}$ | 0.327 | 0.119 | 0.305 | 0.684 | 0.250 |
| $Q^{\mathrm{L}}$ | −0.582 | −0.280 | −0.077 | 0.079 | −0.063 |
| $\Delta Q^{\mathrm{C}}$ | −0.480 | −0.252 | −0.128 | −0.054 | −0.110 |
| $\delta^{\mathrm{MC}}$ | 1.404 | 1.259 | 0.968 | 0.565 | 0.874 |
| $\delta^{\mathrm{MO}}$ | 0.240 | 0.213 | 0.148 | 0.059 | 0.132 |
| $\delta^{\mathrm{CC}}$ | 0.141 | 0.103 | 0.067 | 0.034 | 0.038 |
| $\Delta\delta^{\mathrm{CO}}$ | −0.333 | −0.242 | −0.113 | 0.032 | −0.077 |
| $E_{\mathrm{int}}^{\mathrm{ML}}$ | −0.368 | −0.329 | −0.244 | −0.126 | −0.192 |
| $E_{\mathrm{int}}^{\mathrm{MC}}$ | −0.253 | −0.272 | −0.154 | 0.024 | −0.116 |
| $\Delta E_{\mathrm{int}}^{\mathrm{CO}}$ | 0.441 | 0.302 | 0.206 | 0.149 | 0.177 |
| $V_{\mathrm{cl}}^{\mathrm{ML}}$ | −0.019 | −0.004 | 0.002 | 0.009 | 0.007 |
| $V_{\mathrm{cl}}^{\mathrm{MC}}$ | 0.073 | 0.032 | 0.078 | 0.153 | 0.071 |
| $\Delta V_{\mathrm{cl}}^{\mathrm{CO}}$ | 0.384 | 0.270 | 0.204 | 0.178 | 0.181 |
| $V_{\mathrm{xc}}^{\mathrm{ML}}$ | −0.350 | −0.325 | −0.246 | −0.135 | −0.199 |
| $V_{\mathrm{xc}}^{\mathrm{MC}}$ | −0.327 | −0.304 | −0.232 | −0.129 | −0.187 |
| $\Delta V_{\mathrm{xc}}^{\mathrm{CO}}$ | 0.057 | 0.033 | 0.002 | −0.029 | −0.004 |

[a] Data for the isolated CO species: $d$(CO) = 1.137, $Q^{\mathrm{C}}$ = 1.209, $\delta^{\mathrm{CO}}$ = 1.726, $E_{\mathrm{int}}^{\mathrm{CO}}$ = −1.773, $V_{\mathrm{cl}}^{\mathrm{CO}}$ = −1.310, and $V_{\mathrm{xc}}^{\mathrm{CO}}$ = −0.463. The structure of the table repeats that found in Table 1.

Table 2 summarizes our M06 DFT results in the same set of complexes. We must stress that standard Kohn−Sham (KS) DFT calculations lack a properly defined second order density matrix, so our $V_{\mathrm{xc}}$ and $\delta$ values are approximate, obtained by constructing a Dirac−Fock pseudo-pair density from the KS determinant. There is nevertheless evidence[51] that DFT $\delta$'s do compare reasonably well with wave function based values.

All of our previous arguments apply, mostly unchanged, to the DFT data. This fact seems to support the soundness of these IQA procedures based on Dirac−Fock density matrices constructed from KS determinants. A couple of points deserve mentioning, nevertheless. The DFT free CO molecule description takes into account, even with the above-mentioned approximations, the rather large change in polarization caused by electron correlation. This is seen in the clearly smaller DFT $Q$'s and $V_{\mathrm{cl}}^{\mathrm{CO}}$ interaction and in a noticeably larger $V_{\mathrm{xc}}^{\mathrm{CO}}$. Formation of the complexes gives rise to a decrease in the overall polarity of CO, and both the $Q$ and $V_{\mathrm{cl}}^{\mathrm{CO}}$ values decrease markedly. The DFT free or coordinated CO is clearly less ionic and more covalent than the HF one. As expected, correlation increases back-donation as measured by $\delta^{\mathrm{MO}}$. Notice how in the Cu compound $V_{\mathrm{xc}}^{\mathrm{CO}}$ has now decreased below the free CO value by about 20 kcal/mol. This is consistent with the onset of a nonclassical carbonyl.

## 5. Classical Penta- and Hexacarbonyls

We will summarize in this section our results on the d$^8$ $D_{3h}$ pentacoordinated and octahedral classical carbonyls. Tables 3 and 4 show that most of our comments regarding the M(CO)$_4$ species hold in these compounds.

The pentacarbonyls are characterized by well differentiated axial and equatorial ML bonds. It is well-known that the equatorial link is generally stronger and shorter, but this gets reversed in the Co case. In this latter complex, the overall topological charge of the CO ligands is positive, as in $[\mathrm{Cu(CO)_4}]^+$. It is also known, though nonetheless interesting, that the metal is quite positively charged even in the Mn

***Table 3.*** Geometric, QTAIM Integrated Properties, and IQA Interactions for the d$^8$ $D_{3h}$ $[Mn(CO)_5]^-$, $Fe(CO)_5$, $[Co(CO)_5]^+$, and $Ru(CO)_5$ Complexes[a]

| M | Mn | Fe | Co | Ru |
|---|---|---|---|---|
| $d(MC)_{ax}$ | 1.937 | 2.061 | 2.176 | 2.047 |
| $d(MC)_{eq}$ | 1.822 | 1.875 | 2.251 | 2.043 |
| $\Delta d(CO)_{ax}$ | 0.013 | −0.002 | −0.011 | 0.000 |
| $\Delta d(CO)_{eq}$ | 0.030 | 0.008 | −0.009 | 0.006 |
| $\Delta\nu_{ax}$ | −192 | −11 | 140 | −50 |
| $\Delta\nu_{eq}$ | −360 | −188 | 0 | −113 |
| $Q^M$ | 0.678 | 0.570 | 0.852 | 0.515 |
| $Q^L_{ax}$ | −0.192 | −0.034 | 0.036 | −0.061 |
| $Q^L_{eq}$ | −0.431 | −0.156 | 0.025 | −0.139 |
| $\Delta Q^C_{ax}$ | −0.155 | −0.118 | −0.051 | −0.136 |
| $\Delta Q^C_{eq}$ | −0.369 | −0.232 | −0.050 | −0.209 |
| $\delta^{MC}_{ax}$ | 0.750 | 0.521 | 0.365 | 0.834 |
| $\delta^{MC}_{eq}$ | 1.100 | 0.933 | 0.377 | 0.955 |
| $\delta^{MO}_{ax}$ | 0.110 | 0.059 | 0.030 | 0.106 |
| $\delta^{MO}_{eq}$ | 0.158 | 0.123 | 0.032 | 0.126 |
| $\delta^{CC}_{axeq}$ | 0.105 | 0.069 | 0.038 | 0.055 |
| $\Delta\delta^{CO}_{ax}$ | −0.186 | −0.028 | 0.012 | −0.076 |
| $\Delta\delta^{CO}_{eq}$ | −0.234 | −0.067 | 0.018 | −0.055 |
| $E_{int}^{ML_{ax}}$ | −0.204 | −0.130 | −0.094 | −0.203 |
| $E_{int}^{ML_{eq}}$ | −0.318 | −0.231 | −0.087 | −0.226 |
| $E_{int}^{MC_{ax}}$ | 0.009 | 0.029 | 0.125 | −0.042 |
| $E_{int}^{MC_{eq}}$ | −0.095 | −0.065 | 0.119 | −0.078 |
| $\Delta E_{int}^{CO_{ax}}$ | 0.153 | 0.214 | 0.069 | 0.226 |
| $\Delta E_{int}^{CO_{ax}}$ | 0.309 | 0.344 | 0.059 | 0.317 |
| $V_{cl}^{ML_{ax}}$ | −0.024 | −0.010 | −0.012 | −0.007 |
| $V_{cl}^{ML_{eq}}$ | −0.048 | −0.002 | −0.006 | −0.004 |
| $V_{cl}^{MC_{ax}}$ | 0.178 | 0.143 | 0.204 | 0.144 |
| $V_{cl}^{MC_{eq}}$ | 0.159 | 0.151 | 0.197 | 0.132 |
| $\Delta V_{cl}^{CO_{ax}}$ | 0.124 | 0.225 | 0.084 | 0.230 |
| $\Delta V_{cl}^{CO_{eq}}$ | 0.269 | 0.354 | 0.074 | 0.330 |
| $V_{xc}^{ML_{ax}}$ | −0.180 | −0.120 | −0.081 | −0.195 |
| $V_{xc}^{ML_{eq}}$ | −0.270 | −0.227 | −0.081 | −0.222 |
| $V_{xc}^{MC_{ax}}$ | −0.169 | −0.114 | −0.078 | −0.185 |
| $V_{xc}^{MC_{eq}}$ | −0.254 | −0.215 | −0.078 | −0.210 |
| $\Delta V_{xc}^{CO_{ax}}$ | 0.031 | −0.010 | −0.014 | 0.003 |
| $\Delta V_{xc}^{CO_{eq}}$ | 0.041 | −0.009 | −0.014 | −0.012 |

[a] HF data in atomic units, except distances in Å and frequencies in cm$^{-1}$.

***Table 4.*** Geometric, QTAIM Integrated Properties, and IQA Interactions for the d$^6$ $O_h$ $[Ti(CO)_6]^{2-}$, $[V(CO_6)]^-$, $Cr(CO)_6$, $[Mn(CO)_4]^+$, and $[Fe(CO)_6]^{2+}$ Systems[a]

| M | Ti | V | Cr | Mn | Fe |
|---|---|---|---|---|---|
| $d(MC)$ | 2.036 | 1.986 | 2.012 | 2.159 | 2.225 |
| $\Delta d(CO)$ | 0.041 | 0.023 | 0.005 | −0.008 | −0.016 |
| $\Delta\nu$ | −495 | −288 | −108 | 57 | 148 |
| $Q^M$ | 1.461 | 1.188 | 0.930 | 0.994 | 1.153 |
| $Q^L$ | −0.574 | −0.361 | −0.154 | 0.002 | 0.141 |
| $\Delta Q^C$ | −0.509 | −0.352 | −0.162 | −0.071 | −0.118 |
| $\delta^{MC}$ | 0.622 | 0.739 | 0.680 | 0.444 | 0.374 |
| $\delta^{MO}$ | 0.092 | 0.111 | 0.098 | 0.045 | 0.032 |
| $\delta^{CC}$ | 0.140 | 0.106 | 0.067 | 0.039 | 0.050 |
| $\Delta\delta^{CO}$ | −0.225 | −0.175 | −0.124 | −0.012 | 0.156 |
| $E_{int}^{ML}$ | −0.324 | −0.267 | −0.192 | −0.113 | −0.090 |
| $E_{int}^{MC}$ | 0.071 | 0.060 | 0.067 | 0.137 | 0.156 |
| $\Delta E_{int}^{CO}$ | 0.416 | 0.343 | 0.150 | 0.082 | 0.218 |
| $V_{cl}^{ML}$ | −0.187 | −0.096 | −0.034 | −0.013 | −0.004 |
| $V_{cl}^{MC}$ | 0.200 | 0.221 | 0.216 | 0.232 | 0.239 |
| $\Delta V_{cl}^{CO}$ | 0.376 | 0.319 | 0.135 | 0.093 | 0.272 |
| $V_{xc}^{ML}$ | −0.138 | −0.171 | −0.158 | −0.099 | −0.085 |
| $V_{xc}^{MC}$ | −0.130 | −0.161 | −0.149 | −0.095 | −0.082 |
| $\Delta V_{xc}^{CO}$ | 0.041 | 0.025 | 0.015 | −0.010 | −0.053 |

[a] HF data in atomic units, except distances in Å and frequencies in cm$^{-1}$.

complex. This means that the ligands in the complexes bear a considerable negative charge, larger in the equatorial positions. The values of the average MC delocalization indices are similar to those found in the equivalently charged $T_d$ molecules, and so is the covalent energy associated with the MC bond. There is however a tendency toward stronger (weaker) MC links for the negatively (positively) charged $D_{3h}$ complexes when compared to their equivalently charged $T_d$ counterparts. Together with $\delta^{MO}$, our calculations show that $\pi$-back-bonding is largest in the Mn compound and smallest in the cobalt one. This also justifies the large positive metal charges. Our $V_{xc}$ and $V_{cl}$ values validate that, in general, the axial ML bonds are more ionic in the four $D_{3h}$ cases than the equatorial ones, although in some cases the difference is small. A salient feature of our data is the positive value of these axial $E_{int}^{MC}$'s for all of the systems except $Ru(CO)_5$, so it is the very stabilizing MO interaction which stabilizes the ML link. This is related to the combined effect of the larger positive charges of the M and the axial carbon atoms. Contrarily to the uniform MC electrostatic behavior, there is a clear change in the MC covalency. The axial links in the Mn and Fe moieties are less covalent than the equatorial ones, while the opposite holds in the $Ru(CO)_5$ molecule.

As regards the CO moiety, it is relatively interesting to notice that the equatorial CO distance is larger than the axial one, independently of the M−C distance behavior. This correlates reasonably with delocalization indices and $V_{xc}$ values, and as in the $T_d$ compounds, $\Delta\delta$'s are considerably larger than $\Delta V_{xc}$'s. In the Mn and Fe complexes, the axial carbonyls are slightly more covalent than the equatorial ones, but the situation is clearly reversed for $Ru(CO)_5$. As found in the tetrahedral cases, the total CO electrostatic interaction is related to the repolarization of the carbonyl group, increasing on average with $|\Delta Q^C|$. As with $[Cu(CO)_4]^+$, $\Delta d$ (CO), $Q^L$, $\Delta\delta^{CO}$, $E_{int}^{MC}$, and $\Delta V_{xc}^{CO}$ change sign for $[Co(CO)]^+$. This consistency shows how intimately coupled the changes in the CO ligands are to the ML bonding features.

Finally, it is worthwhile noticing the similarity in the trends of the CO variations upon bonding for equally charged $T_d$ and $D_{3h}$ systems, e.g., $[Co(CO)_4]^-$ and $[Mn(CO)_5]^-$, and the prominent CC delocalizations in the Mn and Fe pentacarbonyls, which again point toward non-negligible multicenter bonding features among the carbonyls with covalent contributions as large as 10 kcal/mol.

Correlation effects may alter significantly the HF geometry of these compounds. For instance, the M06 geometry for $Fe(CO)_5$, one of the systems with more dramatic changes, gives axial and equatorial MC distances of 1.798 and 1.801 Å, respectively, with $\Delta d(CO)$ equal to 0.010 and 0.012 Å in the same order. At this geometry, a HF IQA analysis provides $Q^M = 0.516$, so the total charge transfer has not changed much due to the geometry change, but other quantities depending on the quite shorter MC distances are altered as expected. For instance, $\delta^{MC}_{ax} = 0.829$ and $\delta^{MC}_{eq} = 0.994$. Similarly, $\delta^{MO}_{ax} = 0.106$ and $\delta^{MO}_{eq} = 0.129$. As the ML interactions are regarded, $V_{cl}^{ML_{ax}} = 0.005$ and $V_{cl}^{ML_{eq}} = 0.016$ au, while $V_{xc}^{ML_{ax}} = -0.209$ and $V_{xc}^{ML_{ax}} = -0.239$ au. The decrease in the ML distances thus leads to increased ML interactions, much more important in the axial than in the equatorial link.

The octahedral d$^6$ hexacarbonyls follow similar basic rules. As we move from Ti to Fe, the MC distance increases with the exception of the Ti molecule, and the CO bond length decreases monotonically, getting shorter than in the isolated molecule for both the Mn and Fe complexes, which also show positive $Q^L$ values. Simultaneously, $Q^M$ passes through a minimum in the Cr complex. Notice how, as $Q^{L}$ goes from negative to positive, $\Delta Q^{C}$ becomes decoupled from it. The double negative Ti anion has the largest positive metal charge and the most negatively charged CO species of all the examples examined up to now. However, both $\delta^{MC}$ and $\delta^{MO}$ and their $V_{xc}$ covalent energy counterparts clearly show that back-bonding in these $O_h$ complexes has saturated at the vanadium complex, and that in titanium the approach of the six carbonyls is only possible at a slightly larger MC final distance. This is accompanied by a quite large intercarbonyl delocalization, as measured by $\delta^{CC}$. Another way to look at the same saturation stems from the metal localization index, $\lambda^{Ti} = 18.39$. This parameter grows monotonously up to 23.69 in the Fe hexacarbonyl. Only about 18 electrons (its [Ar] core) are localized in the Ti atomic basin as far as two-center delocalizations are regarded, so all of the valence has been used in bonding to a first approximation. This effect may explain the unexpectedly large TiC distance.

The $O_h$ systems show very large M charges, thus positive total $E_{int}^{MC}$ values independently of the value of $Q^{L}$. This behavior is different from that found in our previous example and makes the MO interaction decisive in accounting for the negative $E_{int}^{ML}$ values and the stability of the complexes. We want to stress that, as we go from the tetra- to the hexacarbonyls, the value of $E_{int}^{ML}$ turns out to be a function of the net charge of the complex and its coordination. For a given total charge of $-2$, $-1$, 0, $+1$, and $+2$, its most negative value, $-0.34$, $-0.31$, $-0.23$, $-0.11$, and $-0.09$ $E_h$, is attained for the $[Fe(CO)_4]^{-2}$, $[Mn(CO)_5]^{-}$, $Fe(CO)_5$, $[Mn(CO)_6]^{+}$, and $[Fe(CO)_6]^{2+}$ complexes, respectively. Thus, the ML interaction is most favorable for middle 3d metals, low coordinations, and negatively charged complexes. However, as the total charge becomes positive, higher coordinations become preferable (notice that we have only one dication in this series).

We also notice that the covalency of the MC interaction decreases on going from anions to cations (as back-bonding arguments suggest) for any coordination, except in $[Ti(CO)_6]^{2-}$, for which we have already suggested a saturation phenomenon. Moreover, the MC $V_{xc}$ decreases with coordination, and as we move from tetra- to hexacarbonyls, the ML link becomes more ionic, as measured by $V_{cl}^{ML}$. This is a very well-known bonding tendency in solid state physics, where larger coordination phases tend to be more ionic and, in fact, a simple consequence of Pauling's rules. Overall, $V_{cl}^{ML}$ is negligibile, except in penta- and hexacoordinated anions, where $V_{xc}^{ML}$ peaks, so almost all of the ML stabilization energy comes from covalent contributions that may be small in cations.

With all the above arguments, the changes in the CO quantities of our $M(CO)_6$ molecules are easily rationalized. As seen in Table 4, both $\Delta d$ (CO) and $\Delta V_{xc}^{CO}$ are negative in the Mn and Fe molecules. Their stronger CO links do also display a positive CO net charge with rather small negative $Q^{O}$ and small $\delta^{MO}$, thus very small back-bonding. As we will explore in the next section, both the change of sign and the decrease in magnitude of the CO polarity are signatures of this behavior.

***Table 5.*** Geometric, QTAIM Integrated Properties, and IQA Interactions for the d$^8$ Square Planar $[Ni(CO)_4]^{2+}$ and $[Pd(CO)_4]^{2+}$ Complexes[a]

| M | Ni | Pd | | Ni | Pd |
|---|---|---|---|---|---|
| $d$(MC) | 2.126 | 2.131 | $Q^{M}$ | 1.490 | 1.104 |
| $\Delta d$(CO) | $-0.019$ | $-0.020$ | $Q^{L}$ | 0.128 | 0.224 |
| $\Delta\nu$ | 172 | 120 | $\Delta Q^{C}$ | $-0.057$ | 0.035 |
| $\delta^{MC}$ | 0.383 | 0.614 | $E_{int}^{ML}$ | $-0.091$ | $-0.101$ |
| $\delta^{MO}$ | 0.026 | 0.050 | $E_{int}^{MC}$ | 0.264 | 0.171 |
| $\Delta\delta^{CO}$ | 0.092 | 0.069 | $\Delta E_{int}^{CO}$ | 0.107 | 0.080 |
| $V_{cl}^{ML}$ | $-0.004$ | 0.042 | $V_{xc}^{ML}$ | $-0.087$ | $-0.143$ |
| $V_{cl}^{MC}$ | 0.348 | 0.309 | $V_{xc}^{MC}$ | $-0.084$ | $-0.138$ |
| $\Delta V_{cl}^{CO}$ | 0.146 | 0.118 | $\Delta V_{xc}^{CO}$ | $-0.038$ | $-0.036$ |

[a] HF data in atomic units, except distances in Å and frequencies in cm$^{-1}$.

## 6. Nonclassical Carbonyls

Let us start by considering the d$^8$ square planar $[Ni(CO)_4]^{2+}$ and $[Pd(CO)_4]^{2+}$ 16-electron complexes. Table 5 summarizes our results. A first look confirms our previous observations: shortening instead of lengthening of $d$ (CO), large metal positive topological charges, positive but small total net charge for the carbonyl ligands, relatively small MC delocalization indices coupled to very small $\delta^{MO}$ or back-bonding, and positive $\Delta\delta^{CO}$. Energetically, we find positive MC interaction energies with large negative $E_{int}^{MO}$ and not too large covalent contributions that point to important MC ionicity, and as far as the CO ligand is regarded, larger covalency than in free CO. This provides an image in which almost all the characteristics of standard carbonyls have been reversed. Notice how, as we move from anionic to cationic species, the overall $V_{cl}^{ML}$ passes from negative values to even destabilizing interactions, here exemplified by the $[Pd(CO)_4]^{2+}$ system. Its only overall stabilizing ML term is $V_{xc}^{ML}$, clearly dominated by the MC contribution.

Since there are a number nonclassical mono- and dicarbonyl cations traditionally considered as nonclassical, we have performed CASSCF calculations on some of them to ascertain the role of at least static correlation on the IQA energetic quantities of these complexes. Our results are gathered in Table 6.

Most of our previous findings apply unchanged. However, caution is necessary when comparing the CASSCF free CO quantities with those found in the complexes, since the limited amount of correlation accounted for does not affect the CO moiety in the same manner when isolated or interacting. Even with this in mind, we clearly see a shortening of the CO bond length in all the cases, positive CO net charges that increase with the coordination index, relatively small MC delocalization indices, and very low $\delta^{MO}$'s, except in the gold complexes, which display a large covalency in the MC link and behave differently, see below. These systems do also show negative $\Delta V_{xc}^{CO}$ values and lowly polarized CO ligands with larger CO total interactions. In the dicarbonyls, $\Delta E_{int}^{CO}$ and even $\Delta V_{cl}^{CO}$ may become negative.

Notice that the total ML interaction energies are not small, peaking at $-109$ kcal/mol for the $Au(CO)^{+}$ molecule, and that the general behavior of the $HCO^{+}$ system parallels rather closely that of the gold compounds.

A comparison among the Cu, Ag, and Au compounds, on one hand, and of mono- and dicarbonyls, on the other, is

***Table 6.*** $HCO^+$, Together with the $d^{10}$ Linear $Cu(CO)^+$, $Ag(CO)^+$, $Au(CO)^+$, $[Cu(CO)_2]^+$, $[Ag(CO)_2]^+$, and $[Au(CO)_2]^+$ Data, in This Order, Calculated at the CASSCF Level with MP2 Geometries, as Described in the Text[a]

| M | H | Cu | Ag | Au | Cu | Ag | Au |
|---|---|---|---|---|---|---|---|
| $d$(MC) | 1.090 | 1.967 | 2.311 | 1.975 | 1.937 | 2.234 | 2.006 |
| $\Delta d$(CO) | −0.015 | −0.007 | −0.006 | −0.006 | −0.006 | −0.005 | −0.007 |
| $\Delta\nu$ | 13 | 95 | 77 | 90 | 107 | 120 | 119 |
| $Q^M$ | 0.414 | 0.887 | 0.889 | 0.782 | 0.786 | 0.768 | 0.645 |
| $Q^L$ | 0.586 | 0.114 | 0.111 | 0.220 | 0.108 | 0.118 | 0.179 |
| $\Delta Q^C$ | 0.281 | −0.033 | −0.011 | 0.043 | 0.030 | 0.113 | 0.151 |
| $\delta^{MC}$ | 0.687 | 0.567 | 0.434 | 0.975 | 0.573 | 0.483 | 0.855 |
| $\delta^{MO}$ | 0.058 | 0.036 | 0.026 | 0.080 | 0.039 | 0.032 | 0.072 |
| $\Delta\delta^{CO}$ | 0.007 | 0.046 | 0.052 | −0.031 | 0.109 | 0.144 | 0.123 |
| $E_{int}^{ML}$ | −0.090 | −0.115 | −0.079 | −0.173 | −0.121 | −0.088 | −0.166 |
| $E_{int}^{MC}$ | 0.034 | 0.077 | 0.093 | 0.004 | 0.069 | 0.087 | 0.015 |
| $\Delta E_{int}^{CO}$ | 0.115 | 0.102 | 0.071 | 0.119 | −0.004 | −0.107 | −0.099 |
| $V_{cl}^{ML}$ | 0.129 | 0.014 | 0.011 | 0.057 | 0.014 | 0.016 | 0.043 |
| $V_{cl}^{MC}$ | 0.247 | 0.202 | 0.180 | 0.226 | 0.200 | 0.188 | 0.216 |
| $\Delta V_{cl}^{CO}$ | 0.146 | 0.118 | 0.086 | 0.132 | 0.011 | −0.092 | −0.084 |
| $V_{xc}^{ML}$ | −0.220 | −0.129 | −0.090 | −0.231 | −0.135 | −0.104 | −0.209 |
| $V_{xc}^{MC}$ | −0.213 | −0.125 | −0.087 | −0.223 | −0.131 | −0.101 | −0.201 |
| $\Delta V_{xc}^{CO}$ | −0.031 | −0.016 | −0.015 | −0.013 | −0.015 | −0.015 | −0.014 |

[a] All data in atomic units, except distances in Å and frequencies in $cm^{-1}$. CASSCF//MP2 data for isolated CO: $d(CO) = 1.147$, $Q^C = 1.188$, $\nu(CO) = 2137\ cm^{-1}$, $\delta^{CO} = 1.384$, $E_{int}^{CO} = -1.711$, $V_{cl}^{CO} = -1.295$, and $V_{xc}^{CO} = -0.416$.

interesting by itself. Our data provide a fresh new perspective to some reported observations. For instance, it has been found[23] that the Cu and Ag dicarbonyls display larger electron and energy densities at the MC bond critical point than those in the monocarbonyls, but that the contrary is true for the Au compounds. This is interpreted in terms of larger/smaller covalent contributions using the empirical correlation between these quantities. Our $V_{xc}^{MC}$ and $V_{xc}^{ML}$ values provide direct support to these claims, being larger in the dicarbonyls except for the Au systems. There is also consensus about significant back-donation in $Au(CO)^+$, small back-donation in $Cu(CO)^+$, and negligible back-donation in $Ag(CO)^+$. Our $\delta^{MO}$ and $V_{xc}^{MO}$ support this statement.

However, a rather constant argument in the literature ascribes a mainly electrostatic character to the ML link in $Ag(CO)^+$, for instance. Our analysis allows us to pinpoint the nature of this claim by using either the atomic or functional group point of view for the CO ligand. The overall AgC interaction energy is positive, thus destabilizing, by 58 kcal/mol, a result coming from a large electrostatic repulsion associated with 113 kcal/mol and a significant covalent interaction of −55 kcal/mol. It is only in this sense that the AgC link is mainly ionic in nature: it is the MO electrostatic attraction associated with an energy of −106 kcal/mol (and a weak covalent term of −2 kcal/mol) which leads to the overall negative AgL interaction of −50 kcal/mol. Shifting to the functional group view of the carbonyl ligand, as is usually done, the picture changes dramatically, for the overall $V_{cl}^{ML}$ is now very small, +7 kcal/mol, and it is not electrostatics, but covalency, which binds the system.

The nature of the stiffening of the CO interaction in nonclassical carbonyls may also be explored with our procedure. We have thus performed equivalent CASSCF calculations to those presented in Table 6 in the $HCO^+$ and $COH^+$ systems for several HC or OH distances and relaxed CO bond lengths. Table 7 gathers the most interesting results.

***Table 7.*** Topological Charges, Sharing Indices, and IQA Properties for the CASSCF Descriptions of $HCO^+$ and $COH^+$ as the HC or OH Distances ($d$) Are Varied with Respect to Their Respective Equilibrium Values: $\Delta d = d - d_{eq}$[a]

| | | | $HCO^+$ | | | |
|---|---|---|---|---|---|---|
| $\Delta d$ | −0.2 | −0.1 | 0.0 | 0.1 | 0.2 | 0.3 |
| $\Delta d$(CO) | −0.002 | −0.001 | 0.000 | 0.004 | 0.005 | 0.005 |
| $Q^L$ | 0.699 | 0.574 | 0.521 | 0.554 | 0.529 | 0.511 |
| $\Delta Q^C$ | 0.418 | 0.302 | 0.257 | 0.268 | 0.248 | 0.234 |
| $Q^H$ | −0.301 | 0.426 | 0.477 | 0.443 | 0.469 | 0.488 |
| $\delta^{HC}$ | 0.884 | 0.791 | 0.744 | 0.650 | 0.615 | 0.584 |
| $\Delta\delta^{CO}$ | 0.042 | 0.054 | 0.059 | 0.011 | 0.010 | 0.010 |
| $\Delta E_{int}^{CO}$ | 0.048 | 0.068 | 0.067 | 0.086 | 0.082 | 0.079 |
| $\Delta V_{cl}^{CO}$ | 0.083 | 0.104 | 0.103 | 0.121 | 0.116 | 0.112 |
| $\Delta V_{xc}^{CO}$ | −0.035 | −0.036 | −0.036 | −0.035 | −0.034 | −0.033 |
| | | | $COH^+$ | | | |
| $\Delta d$(CO) | −0.001 | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 |
| $Q^L$ | 0.410 | 0.324 | 0.270 | 0.232 | 0.202 | 0.190 |
| $\Delta Q^C$ | 0.274 | 0.263 | 0.252 | 0.242 | 0.231 | 0.220 |
| $Q^H$ | 0.590 | 0.675 | 0.730 | 0.770 | 0.800 | 0.811 |
| $\delta^{HO}$ | 0.595 | 0.484 | 0.403 | 0.343 | 0.297 | 0.276 |
| $\Delta\delta^{CO}$ | −0.255 | −0.269 | −0.218 | −0.200 | −0.182 | −0.168 |
| $\Delta E_{int}^{CO}$ | 0.104 | 0.079 | 0.067 | 0.062 | 0.059 | 0.061 |
| $\Delta V_{cl}^{CO}$ | 0.022 | 0.000 | −0.008 | −0.009 | −0.008 | −0.002 |
| $\Delta V_{xc}^{CO}$ | 0.083 | 0.079 | 0.075 | 0.071 | 0.067 | 0.063 |

[a] CO distances for the optimum $HCO^+$ and $COH^+$ molecules are 1.119 and 1.166 Å and should be compared to the isolated CASSCF CO one, 1.150 Å. All data in atomic units, except distances in Å.

First, the effect of the electric field imposed by the H proton is clear, and the polarization of the CO group is quite different in both cases. The proton is shielded much more efficiently in the $HCO^+$ case, as the $Q^H$ value shows, even though it approaches the CO moiety toward the positively charged end. This already points toward covalency as an important factor in accounting for the stiffening phenomenon, more important for a HC interaction than for a HO one due to the smaller electronegativity difference between the atoms in the first case. Notice that the CO distance decreases from the isolated CO molecule in the $HCO^+$ case, but increases in the $COH^+$ one.

The origin of this very different behavior may be traced to the HC and HO delocalization indices in Table 7, which have a huge impact on $\delta^{CO}$. The latter is much smaller in the $COH^+$ system, and this propagates the covalent energies, which are more stabilizing than those in the free CO molecule for the $HCO^+$ case and considerably less stabilizing in $COH^+$. Notice how, from all the data contained in the table, the behavior of $\Delta\delta^{CO}$ and $\Delta V_{xc}^{CO}$ stands out. As found before, $\delta$'s are very sensitive indicators of the CO behavior. The CO stiffening in $HCO^+$ is thus a covalent effect, which as our examples show, and in agreement with Lupinetti and co-workers,[23] is triggered by the particular polarization pattern generated by an $RCO^+$ arrangement, characterized by a smaller CO charge separation than in $COR^+$. We want to stress that, as found in other IQA studies,[36] bond lengths and probably force constants are strongly correlated to the $V_{xc}$ component of the interaction energy. We can see that the CO electrostatic contribution is clearly more stabilizing in the $COH^+$ arrangement, and that even the total $E_{int}^{CO}$ may exceed that found in $HCO^+$. However, the general $d^{-1}$ distance dependence of $V_{cl}$ makes this term vary more slowly than $V_{xc}$, which changes exponentially with the distance of

***Table 8.*** Topological Charges and Sharing Indices for Several DFT Descriptions of the $[Ag(CO)_2]^+$ and $[Au(CO)_2]^+$ Systems, Together with HF Data, All of Them at the MP2 Geometry[a]

| | B3LYP | BLYP | BLYP-LC | M06-l | M06 | M06-HF | HF |
|---|---|---|---|---|---|---|---|
| $Q^{Ag}$ | 0.711 | 0.695 | 0.701 | 0.728 | 0.706 | 0.739 | 0.770 |
| $Q^{C}$ | 1.189 | 1.147 | 1.176 | 1.187 | 1.208 | 1.237 | 1.338 |
| $Q^{O}$ | −1.044 | −0.995 | −1.027 | −1.050 | −1.060 | −1.107 | −1.223 |
| $\delta^{AgC}$ | 0.576 | 0.610 | 0.579 | 0.573 | 0.573 | 0.510 | 0.482 |
| $\delta^{AgO}$ | 0.063 | 0.074 | 0.059 | 0.069 | 0.060 | 0.046 | 0.042 |
| $\delta^{CC}$ | 0.013 | 0.017 | 0.011 | 0.013 | 0.012 | 0.006 | 0.005 |
| $\delta^{CO}$ | 1.795 | 1.840 | 1.815 | 1.782 | 1.769 | 1.748 | 1.599 |
| $Q^{Au}$ | 0.624 | 0.622 | 0.612 | 0.650 | 0.638 | 0.596 | 0.652 |
| $Q^{C}$ | 1.200 | 1.152 | 1.190 | 1.193 | 1.209 | 1.272 | 1.371 |
| $Q^{O}$ | −1.012 | −0.962 | −0.996 | −1.017 | −1.028 | −1.070 | −1.196 |
| $\delta^{AuC}$ | 0.948 | 0.979 | 0.941 | 0.953 | 0.942 | 0.880 | 0.849 |
| $\delta^{AuO}$ | 0.119 | 0.133 | 0.111 | 0.127 | 0.114 | 0.092 | 0.085 |
| $\delta^{CC}$ | 0.047 | 0.055 | 0.042 | 0.047 | 0.040 | 0.034 | 0.027 |
| $\delta^{CO}$ | 1.748 | 1.791 | 1.772 | 1.737 | 1.732 | 1.709 | 1.559 |

[a] All data in atomic units.

the two centers. Larger classical terms at the expense of covalent contributions, as found in this case, usually lead to larger bond lengths and smaller force constants. Similar polarizations are obtained if the H atom is substituted by a positive point charge.

A comment on the values of $\delta$ upon inclusion of electron correlation is needed. It is now known that electron correlation tends to localize electrons in atomic basins beyond the HF independent electron model, so the number of effective pairs of electrons participating in bonding decreases, sometimes dramatically.[35] In this work, for instance, free $\delta^{CO}$ changes from 1.51 (HF) to 1.38 (CASSCF) on including our limited amount of correlation. Simple density functional calculations provide a much larger value, about 1.7, as seen in the caption of Table 2. This is due to the pseudo-single-determinant structure of the KS description, and to the smaller CO bond length predicted at this level of theory. A definitive value for CO is lacking, but $\delta$ changes are well reproduced by any description, as we are seeing.

We have also performed a comparison of the IQA descriptions for these linear carbonyls when several levels of theory are employed, including the approximate DFT descriptions previously described. Table 8 contains a survey of topological charges and sharing indices for the $[Ag(CO)_2]^+$ and $[Au(CO)_2]^+$ systems at the DFT level, with several density functionals to be compared with the CASSCF data contained in Table 6. Notice that introducing HF exchange increases, in general, the topological net charge of C and O and approaches the CASSCF values, which include a very limited amount of correlation (compare with the HF column). Any functional provides noticeably larger delocalization indices than our CASSCF results, with $\delta^{CO}$ close to 1.8. A comment on this large value has already been made in this paper. The DFT data also confirm how electron correlation increases back-donation, as measured by $\delta^{MO}$, in agreement with general knowledge.

## 7. Discussion and Conclusions

The results explored in the previous sections show that the IQA view provides a real space image of bonding in simple metal carbonyls that is compatible with existing knowledge. Overall, IQA interactions show several general features that deserve further comment.

Most, if not all, of the IQA quantities show a very clear dependence on the total net charge and coordination of the complexes examined. As far as the ML interaction is concerned, for instance, Tables 1−6 show that $V_{cl}^{ML}$ is basically controlled by the stoichiometry. As we move from mono- to hexacoordinated complexes, there is a clear tendency for the ML classical interaction to progress from destabilizing to stabilizing values. For a given stoichiometry, it evolves toward increasing stabilization as the net charge of the complex passes from positive to negative values. These trends are a result of the expected increase that $Q^M$ experiences as both the coordination index and the net charge of the complex grow. ML covalency, measured by $V_{xc}^{ML}$, is again dependent on both parameters, but here a more subtle balance takes place. On one hand, larger coordinations saturate the metal binding ability, so $V_{xc}^{ML}$ per ML link tends to decrease on going from mono- to hexacoordinated molecules. On the other hand, extra electrons minimize this saturation tendency, so the largest ML covalent energies appear on the negatively charged tetrahedral carbonyls. As a result of this interplay, except in the octahedral compounds, the total $E_{int}^{ML}$ is dominated by covalency, and its largest values (which are always stabilizing) again occur for negatively charged tetrahedral, pentacoordinated, and octahedral systems, where large $\pi$-back-bonding exists. Similarly, the smallest total ML interactions are found in some positively charged complexes.

The change in the local properties of the CO moiety upon bonding has played a dominant role in the chemistry of metal carbonyls. We have found almost a full match between traditional thinking and the IQA relaxation quantities for CO. Figure 1 shows the change in the CO interaction parameters upon bonding. There, it is clear that the total CO interaction becomes clearly destabilized upon formation of the complexes, except in the Ag and Au dicarbonyls, and that this effect is overwhelmingly dominated by the change suffered by the electrostatic component, $\Delta V_{cl}^{CO}$, which may be obtained by subtracting the much smaller xc component of the lower diagram. Polarization of the CO charge density upon coordination (which includes a charge transfer component) becomes the dominant energetic effect in CO. This conclusion is not affected much by the inclusion of $\Delta E_{self}$ for the C and O atoms, which is almost always negative, so the total CO deformation is generally positive, with minima appearing when $Q^M$ is smallest for each stoichiometry. The change in the CO covalency, shown in the lower diagram of Figure 1, has been discussed in a coordination-like manner in previous sections. Here, it serves us to show that it is the change in the CO covalency which correlates with the CO distance and stretching frequency. So, despite the large electrostatic contributions due to M↔L charge transfer and polarization, it is electron sharing, i.e., electronic effects, which determines the basic signature of ML bonding.

Finally, we will consider an interesting correlation shown by our data.[33,34,42] As already commented upon, the value of $\delta^{MO}$, expected to include negligible $\sigma$ contributions, has been proposed as a measure of the intensity of $\pi$-back-

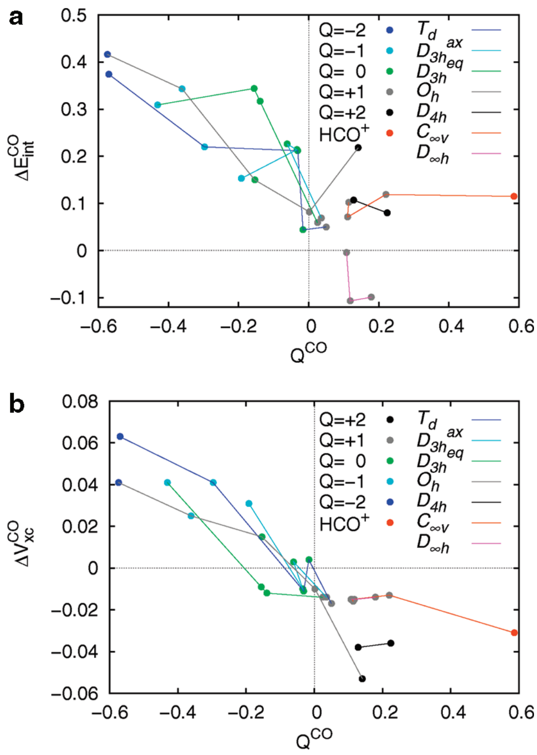


***Figure 1.*** Comparison of the behavior of $\Delta V_{cl}^{CO}$ (upper diagram) and $\Delta V_{xc}^{CO}$ (lower diagram) versus the ligand topological charge ($Q^{CO}$) for the carbonyls examined in this work. Systems have been gathered by symmetry and stoichiometry, and a color code has been added to identify the total net charge of the complex.

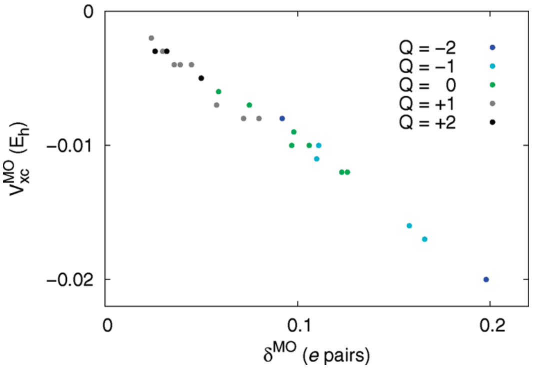


***Figure 2.*** Correlation between $V_{xc}^{MO}$ and $\delta^{MO}$ for the systems studied in this work. The total net charge of the complexes is indicated by the same color code used in Figure 1.

bonding in these systems.[5,25] We similarly expect that $V_{xc}^{MO}$ will extract its energetic content. Figure 2 shows a nice linear correlation between these two quantities. The largest and smallest covalent energies and delocalization indices correspond to negatively and positively charged complexes, respectively. We must again stress that the small value of the energies involved, smaller than 15 kcal/mol, does not measure the energetic intensity of $\pi$-back-donation. It will always be the $V_{xc}^{MC}$ component, mixed with $\sigma$ donation, that dominates. Anyway, our data show that $\delta^{MO}$, much easier to compute than $V_{xc}^{MO}$, may be used safely to measure $\pi$-back-donation. We will corroborate this in a forthcoming paper in which $V_{xc}$'s and $\delta$'s will be decomposed using DAFHs.

To conclude, we have applied the interacting quantum atoms approach (IQA) within the QTAIM to explore chemical bonding in real space in simple transition metal carbonyls. This is the first time that such an energetic viewpoint is provided for TM complexes, thanks to recent protocols devised to deal with effective core potentials in the IQA scheme.[43]

We have explored several classical and nonclassical compounds, with different stereochemistries, and at different levels of theory. As a general conclusion, the bonding image provided by our procedure does not change qualitatively with the inclusion of electron correlation, as noticed by other authors, and is consistent with previous knowledge. It however sheds light on some issues by providing an orbital invariant energetic description of the several metal−carbonyl interactions.

The topological charges of the metal have been found to be positive in all the systems examined, even in dianions. This is known, but clearly shows that formal oxidation states must be taken with care. As expected, $\pi$-back-donation in the standard DCD model accounts reasonably well for our findings, being cleanly related to the amount of electron sharing between the metal and the oxygen atom of each carbonyl group. Negative or neutral complexes show the traditional CO bond length elongation, accompanied by a decrease in both the CO bond order and the covalent energy of the CO bond. Interestingly, our analyses point toward non-negligible multicenter character of the ML bonds, and several carbonyl groups may be involved in it when back-bonding is prominent, as revealed by unexpectedly large intercarbonyl CC delocalization indices. This issue needs to be further explored and will be the subject of further studies.

Covalency dominates the ML link and decreases on going from anions to cations for a given coordination and decreases as coordination increases. Simultaneously, the metal to ligand electrostatic interaction becomes more stabilizing on going from tetra- to hexacarbonyls, mimicking well-known solid state physics behaviors. $V_{cl}^{ML}$, although smaller than $V_{xc}^{ML}$, covers a rather full spectrum, from rather stabilizing (up to −120 kcal/mol) in negatively charged hexacarbonyls, passing through practically electroneutral in most tetrahedral compounds, to clearly destabilizing in some nonclassical systems. It is particularly interesting that the latter, which were originally thought to be mainly bonded by electrostatic forces, tend to be those systems which would not be stable without covalent contributions.

The IQA description of nonclassical carbonyls recovers many of the features already reported using other bonding analyses, showing that these features are rather robust, for instance, the larger covalency of the $[Ag(CO)_2]^+$ system with respect to the $Ag(CO)^+$ one and the reverse behavior of the gold cases. The stiffening of the CO bond is neatly revealed by larger covalent contributions for the CO interaction than those found in the isolated carbon monoxide molecule, even though the energetic changes derived from CO repolarization (as measured by $\Delta V_{cl}^{CO}$) are much larger. Stiffening is thus shown not to be a simple consequence of the electrostatic field imposed by the metal positive net charge, but of the complex reorganization of the CO moiety induced by it.

The energetic perspective provided by IQA may thus be added to the toolkit that the modern theory of chemical bonding in real space provides in transition metal chemistry. By adding interaction energy terms to other well tested indices, it may provide new insights into the field.

**Acknowledgment.** D.T. thanks the University of Milan and the Italian Ministry of Research (MIUR) for a Ph.D. grant that allowed his stay at Oviedo. M.A.B., E.F., and A.M.P. acknowledge financial support from the Spanish MICINN, Project Nos. CTQ2006-02976 and CTQ2009-08376, the European Union FEDER funds, the MALTA-Consolider program (CSD2007-00045), and the FICYT (IB09-019). P.M. thanks the Swiss National Science Foundation for support (project 200021_125313).

## References

(1) Dewar, M. *Bull. Soc. Chim. Fr.* **1951**, *18*, C79.

(2) Chatt, J.; Duncanson, L. A. *J. Chem. Soc.* **1953**, 2929.

(3) Ziegler, T.; Rauk, A. *Inorg. Chem.* **1979**, *18*, 1755.

(4) Weinhold, F.; Landis, C. *Valency and Bonding. A Natural Bond Orbital Donor-Acceptor Perspective*; Cambridge Univ. Press: Cambridge, U. K., 2005.

(5) Macchi, P.; Sironi, A. *Coord. Chem. Rev.* **2003**, *238−239*, 383.

(6) Cortés-Guzmán, F.; Bader, R. F. W. *Coord. Chem. Rev.* **2005**, *105*, 3911.

(7) Pilme, J.; Silvi, B.; Alikhani, M. E. *J. Phys. Chem. A* **2003**, *107*, 4506.

(8) Matito, E.; Sola, M. *Coord. Chem. Rev.* **2009**, *253*, 647.

(9) Dapprich, S.; Frenking, F. *J. Phys. Chem.* **1995**, *99*, 9352.

(10) Davidson, E. R.; Kunze, K. L.; Machado, F. C. B.; Chakravorty, S. *Acc. Chem. Res.* **1993**, *26*, 628.

(11) Pyykkö, P. *Chem. Rev.* **1988**, *88*, 563.

(12) Johnson, J. B.; Klemperer, W. G. *J. Am. Chem. Soc.* **1977**, *99*, 713.

(13) D'Amico, K. L. D.; Trenary, M.; Shin, N. D.; Solmon, E. I.; McFeely, F. R. *J. Am. Chem. Soc.* **1982**, *104*, 5102.

(14) Cotton, F. A.; Wing, R. M. *Inorg. Chem.* **1965**, *4*, 314.

(15) Cotton, F. A. *Inorg. Chem.* **1963**, *3*, 702.

(16) Hall, M. B.; Fensk, R. F. *Inorg. Chem.* **1972**, *11*, 1619.

(17) Hurlbut, P. K.; Rack, J. J.; Luck, J. S.; Dec, S. F.; Webb, J. D.; Anderson, O. P.; Strauss, S. H. *J. Am. Chem. Soc.* **1994**, *116*, 10003.

(18) Souma, Y.; Sano, H. *J. Org. Chem.* **1973**, *38*, 3633.

(19) Bach, C.; Wilner, H. *Angew. Chem., Int. Ed.* **1996**, *108*, 2104.

(20) Sierraalta, A.; Frenking, G. *Theor. Chim. Acta* **1997**, *95*, 1.

(21) Szilagyi, R. K.; Frenking, F. *Organometallics* **1997**, *16*, 4807.

(22) Goldman, A. S.; Krogh-Jespersen, K. *J. Am. Chem. Soc.* **1996**, *118*, 12159.

(23) Lupinetti, A. J.; Fau, S.; Frenking, F.; Strauss, S. H. *J. Phys. Chem.* **1997**, *101*, 9551.

(24) Ehlers, A. W.; Dapprich, S.; Vyboishchikov, S. F.; Frenking, F. *Organometallics* **1996**, *15*, 105.

(25) Macchi, P.; Garlaschelli, L.; Sironi, A. *J. Am. Chem. Soc.* **2002**, *124*, 14173.

(26) Ponec, R.; Lendvay, G.; Chaves, J. *J. Comput. Chem.* **2008**, *29*, 1387.

(27) Ponec, R. *J. Math. Chem.* **1997**, *21*, 323.

(28) Ponec, R. *J. Math. Chem.* **1998**, *23*, 85.

(29) Scherer, W.; Eickerling, D.; Shorokhov, D.; Gullo, G. S.; McGrady, M.; Sirsch, P. *New J. Chem.* **2006**, *30*, 309.

(30) Martín Pendás, A.; Blanco, M. A.; Francisco, E. *J. Chem. Phys.* **2004**, *120*, 4581.

(31) Martín Pendás, A.; Francisco, E.; Blanco, M. A. *J. Comput. Chem.* **2004**, *26*, 344.

(32) Blanco, M. A.; Martín Pendás, A.; Francisco, E. *J. Chem. Theory Comput.* **2005**, *1*, 1096.

(33) Francisco, E.; Martín Pendás, A.; Blanco, M. A. *J. Chem. Theory Comput.* **2006**, *2*, 90.

(34) Martín Pendás, A.; Blanco, M. A.; Francisco, E. *J. Comput. Chem.* **2007**, *28*, 161.

(35) Martín Pendás, A.; Francisco, E.; Blanco, M. A. *J. Phys. Chem. A* **2006**, *110*, 12864.

(36) Martín Pendás, A.; Blanco, M. A.; Francisco, E. *J. Chem. Phys.* **2006**, *125*, 184112.

(37) Martín Pendás, A.; Blanco, M. A.; Francisco, E. *J. Comput. Chem.* **2009**, *30*, 98.

(38) Martín Pendás, A.; Francisco, E.; Blanco, M. A.; Gatti, C. *Chem.—Eur. J.* **2007**, *13*, 9362.

(39) Francisco, E.; Martín Pendás, A.; Blanco, M. A. *J. Chem. Phys.* **2007**, *126*, 094102.

(40) Martín Pendás, A.; Francisco, E.; Blanco, M. A. *J. Phys. Chem. A* **2007**, *111*, 1084.

(41) Martín Pendás, A.; Francisco, E.; Blanco, M. A. *J. Chem. Phys.* **2007**, *127*, 144103.

(42) Martín Pendás, A.; Francisco, E.; Blanco, M. A. *Phys. Chem. Chem. Phys.* **2007**, *9*, 1087.

(43) Tiana, D.; Francisco, E.; Blanco, M. A.; Martín Pendás, A. *J. Phys. Chem. A* **2009**, *113*, 7963.

(44) Bader, R. F. W. *Atoms in Molecules*; Oxford University Press: Oxford, U. K., 1990.

(45) McWeeny, R. *Methods of Molecular Quantum Mechanics*, 2nd ed.; Academic Press: London, 1992; Chapter 14.

(46) Sherwood, D. E.; Hall, M. B. *Inorg. Chem.* **1983**, *22*, 93.

(47) Barnes, A.; Ros, A.; Bauschlicher, C. W., Jr. *J. Chem. Phys.* **1990**, *93*, 609.

(48) Lupinetti, A. J.; Jonas, V.; Thiek, W.; Strauss, S. H.; Frenking, F. *Chem.—Eur. J.* **1999**, *5*, 2573.

(49) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347.

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

(51) Wang, Y. G.; Werstiuk, N. H. *J. Comput. Chem.* **2003**, *24*, 379.

CT9006629

JCTC Journal of Chemical Theory and Computation

# MPI/OpenMP Hybrid Parallel Algorithm for Hartree−Fock Calculations

Kazuya Ishimura,* Kei Kuramoto, Yasuhiro Ikuta, and Shi-aki Hyodo

*Materials Fundamental Research Division, Toyota Central Research & Development Laboratories, Inc., Nagakute, Aichi 480-1192, Japan*

Received December 22, 2009

**Abstract:** A new message passing interface/open multiprocessing (MPI/OpenMP) hybrid parallel algorithm of the Hartree−Fock calculation is developed and implemented into the quantum chemistry program package GAMESS. In the algorithm, internode distribution is performed by MPI and intranode parallelization by OpenMP. It is applied to a $TiO_2$ cluster ($Ti_{35}O_{70}$, 6-31G, 1635 basis functions) and to insulin ($C_{257}H_{381}N_{65}O_{77}S_6^{2-}$, STO-3G, 2430 basis functions) using the Cray XT5 supercomputer (quad-core Opteron 2.3 GHz, 2048 CPU cores). The speed-ups of the whole calculations, including the initial guess generation, are 1238 and 745 using 2048 CPU cores for the $TiO_2$ cluster and insulin, respectively. Hartree−Fock calculations with hundreds or thousands of CPU cores are now practical.

## 1. Introduction

Quantum chemistry calculations have played important roles for analysis and prediction of chemical reactions, molecular spectra, and so on. Calculated molecular systems are becoming larger and larger, and large and complicated systems, such as nanomaterials, self-assembled materials, and biological materials, are now challenging targets. Such systems could introduce unique properties owing to the steric effects with bulky substituents or to a lot of noncovalent interactions, such as dispersions and hydrogen bonds, although the high computational cost is required. Because the size is one of the important factors, calculations of realistic systems are necessary. Several approaches have been proposed to make these calculations possible for large molecules. These approaches could be categorized into mostly two different ways, the reduction of computational costs and the acceleration of calculations. While most of the former are based on modeling by dividing large systems into several or many parts, the latter uses high-performance accelerators or a number of computers.

One of the approaches is the quantum mechanical/molecular mechanical (QM/MM) method[1,2] in which a system is divided into a small reaction center as the QM region and the other as the MM region. The ONIOM method[3,4] divides a molecular system into two or three layers which are treated with different methods and basis sets. These methods can reduce the computational costs by dividing a large molecular system and calculating the important part with a high-level method and the others with a low-level method. A different approach to divide a system into small fragments, called the fragment molecular orbital (FMO) method,[5,6] is also developed, which is especially suitable to proteins. The total computational cost can be reduced because the cost of each fragment is low. However, these methods include approximations or cut-offs, and their accuracies have been discussed by changing computational models or by the combination of computational methods.[7,8]

On the other hand, parallel calculations using central processing units (CPUs), graphics processing units (GPUs), and accelerators are now common because the performance improvement of single CPU cores almost stopped due to heat and power problems. GPUs[9–14] and accelerators[15] are applied to Hartree−Fock (HF), density functional theory (DFT), and quantum Monte Carlo (QMC) calculations. Many efforts have been made for effective GPU computation, for instance, the use of both single- and double-precision values to minimize numerical errors, and the sorting of bra- and ket-pairs for two-electron repulsion integrals (2-ERIs) to utilize the Schwartz screening[16] and to calculate ERIs of same type together.

We can now use hundreds or thousands of CPU cores of supercomputers and PC clusters, and the number of available

* Corresponding author. E-mail: e1502@mosk.tytlabs.co.jp.

10.1021/ct100083w © 2010 American Chemical Society
Published on Web 03/23/2010

CPU cores is increasing. Recent supercomputers have more than 100 000 CPU cores.[17] The parallel efficiency has to be raised for such computer systems by improving the load-balancing and parallelization ratio. Furthermore, large memory space is required for electron correlation calculations, such as perturbation and coupled cluster theories. If only the message passing interface (MPI), which is the most commonly used method, is applied to parallel calculations, the available memory space per process becomes small, because each CPU core allocates each array.

One of the solutions for these problems of CPU parallel calculations is global memory access models such as global arrays[18,19] and distributed data interface (DDI)[20] in which data of other nodes can be accessed through network communication. Another solution is to introduce open multiprocessing (OpenMP), which is a method of intranode parallelization. This supports dynamic load balancing and memory array sharing within a node. OpenMP parallelization has been applied to ab initio calculations[21–23] for multicore/multisocket shared-memory processors. Moreover, the combination of MPI and OpenMP can improve the parallel efficiency and available memory size, in which internode parallelization is performed by MPI and intranode parallelization by OpenMP. The MPI/OpenMP hybrid parallelization makes large molecular calculations using a high level of theory possible without errors caused by approximate models. The hybrid parallelization has been introduced into various calculations, QMC,[24] tight-binding,[25] four-atom QM,[26] and Car−Parrinello molecular dynamics.[27] A similar approach, the combination of Linda and OpenMP, has been implemented into the Gaussian program package.[28] Another approach, the combination of MPI and Pthreads,[29] has been employed to a four-index transformation of 2-ERIs for electron correlation methods in the massively parallel program suite MPQC.[30] Communication bottlenecks are avoided by creating computation and communication threads and overlapping them.

The distribution of the 2-ERI calculation is one of the most important steps to apply the hybrid parallelization to quantum chemistry calculations. In this study, the MPI/OpenMP hybrid parallelization technique is introduced to the HF calculation, which is the basic theory of quantum chemistry. For self-consistent field (SCF) calculations, efficient distributed data parallel algorithms[31,32] have been proposed for distributed memory platforms. Our target is to calculate nanosize molecules which consist of hundreds of atoms using hundreds or thousands of CPU cores. For such computer systems, network communication could be a critical bottleneck. Therefore, we employ replicated data parallelization to reduce and control communication. Furthermore, the initial guess calculation is also parallelized and accelerated to improve the parallel efficiency of the whole calculation. On the basis of the HF parallelization, algorithms of DFT and electron correlation theories will be developed for large molecules, and most calculations based on the QM method will become faster.

```
!$OMP parallel do schedule(dynamic,1) reduction(+:Fock)
  do M= nshell, 1, -1
    do N = 1, M
      MN = M(M- 1) + N
      Λstart = mod(MN + mpi_rank, nproc) + 1
      do Λ = Λstart, M, nproc
        do Σ = 1, Λ
          Schwartz screening
          calculate (μν|λσ) and add into Fock matrix
        end do
      end do
    end do
  end do
!$OMP end parallel do
  Call MPI_AllReduce(Fock)
```

***Figure 1.*** MPI/OpenMP hybrid parallel algorithm for two-electron integral generation.

## 2. Algorithm

The HF calculation mainly consists of the initial guess, Fock matrix generation, and new MO coefficient generation from the Fock matrix. The most time-consuming step is the Fock matrix **F** generation from the 2-ERI $(\mu\nu|\lambda\sigma)$ and the density matrix **D**,

$$(\mu\nu|\lambda\sigma) = \int \frac{\chi_\mu(r_1)\chi_\nu(r_1)\chi_\lambda(r_2)\chi_\sigma(r_2)}{r_1 - r_2} dr_1 dr_2 \qquad (1)$$

$$\mathbf{F}_{\mu\nu} = \mathbf{H}_{\mu\nu} + \sum_{\lambda\sigma} \mathbf{D}_{\lambda\sigma}\{2(\mu\nu|\lambda\sigma) - (\mu\lambda|\nu\sigma)\} \qquad (2)$$

where $\chi$ and **H** denote the contracted basis function and the one-electron integral matrix, respectively.

A new parallel algorithm for the 2-ERI generation is shown in Figure 1. 2-ERIs are generated in the quadruple loop of basis shells, M, N, Λ, and Σ. Indices of the first loop are distributed in intranode by OpenMP and indices of the third loop are distributed in internode by MPI ranks. At the beginning of a calculation, MPI ranks are set to each process.

The OpenMP parallelization is performed at the outermost loop to reduce OpenMP overheads, such as thread generation and data copy. The dynamical distribution and the order of the loop index from a large to a small task are applied to obtain better load balancing. The Fock matrix is allocated by each thread before the quadruple loop and accumulated to the master thread after the loop. Other valuables in the loop are distinguished into shared and private ones. Basis functions, coordinates, molecular orbital, and density matrices are shared with all threads in a process, and blocks of 2-ERIs and intermediate valuables are not shared. Therefore, all large arrays except for the Fock matrix are shared in a node, and data synchronization of threads is not needed during the quadruple loop. Shared valuables are set as common or module ones, and private valuables are set as subroutine arguments of the Fortran language.

The third loop is chosen for the parallelization by MPI ranks from the balance between the parallel granularity and the cost, because the granularity of divided tasks becomes small in an inner loop and the cost of distribution, such as counters, becomes large in an inner loop. To reduce the cost of distribution, an if clause is not used, and the counter is located in the second loop. After the calculation finishes in all processes, the Fock matrix is accumulated, and all processes have the full Fock matrix.

In the initial guess calculation, the extended Hückel (EH) calculation is performed, and the obtained orbitals are projected from the basis for the EH calculation to the basis for the HF calculation. The following formula[33] including many matrix−matrix multiplications is used to make parallelization easy and to reduce the number of floating operations, which is used in the parallel quantum solutions (PQS) program suite:[34]

$$C_1 = S_{11}^{-1} S_{12} C_2 \{ C_2^t S_{12}^t S_{11}^{-1} S_{12} C_2 \}^{-1/2} \tag{3}$$

where $C_1$ is the initial guess molecular orbital (MO) coefficient for the HF calculation, $C_2$ is the EH MO coefficient, $S_{11}$ is the overlap integrals of the basis for the HF calculation, and $S_{12}$ is the overlap integrals of the bases between the HF and the EH calculations.

The Fock matrix diagonalization is performed only at the first and last iterations. During HF iterations, an approximate second-order self-consistent field (SOSCF) method[35,36] is applied to reduce the number of matrix diagonalizations because linear algebra package (LAPACK) routines are parallelized only in intranode.

Intranode parallelization of matrix−matrix multiplications and diagonalization is achieved using thread-parallel basic linear algebra subprograms (BLAS) and LAPACK libraries. In the calculation of matrix−matrix multiplications, columns of the matrix on the right side are distributed to each process, and the result of each process is collected to the master process. The diagonalization calculation is performed only in the master process, and then obtained molecular energies and orbitals are distributed to all processes.

## 3. Results and Discussion

The algorithm was implemented into the quantum chemistry program package GAMESS[37] version April 2008. Benchmark calculations on Cray XT5 (Opteron 2.4 GHz Shanghai core, 512 KB L2-Cache and 6 MB shared L3-Cache, 8 CPU cores per node) 2048 CPU cores were performed using a $TiO_2$ cluster ($Ti_{35}O_{70}$, 6-31G, 1645 basis functions) as a complicated system including d electrons and insulin ($C_{257}H_{381}N_{65}O_{77}S_6^{2-}$ (PDB code: 1HIU),[38] STO-3G, 2430 basis functions) as a simple system in which only s and p basis functions are used. The computer has a PGI Fortran compiler-8.0.2, LIBSCI-10.3.3.5 as a BLAS and LAPACK library, and XT-mpt-3.1.2.1 based on MPICH2−1.0.6p1 as an MPI library. The divide and conquer method[39] is used for diagonalization calculations. The threshold of the density matrix convergence is set to be $1.0 \times 10^{-4}$. The numbers of SCF cycles for the $TiO_2$ cluster and insulin are 30 and 45, respectively.

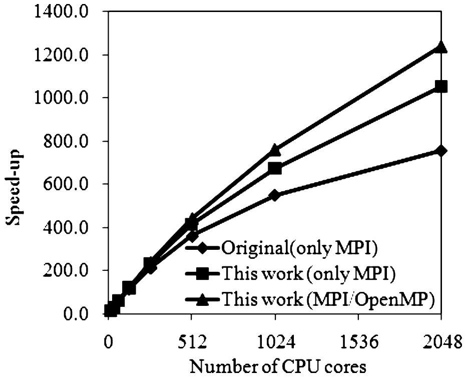


***Figure 2.*** Speed-up of whole Hartree−Fock calculation for $TiO_2$ cluster.

Three kinds of parallel calculations were performed using the following programs: the original GAMESS program using only MPI, the developed programs using only MPI, and MPI for internode and OpenMP for intranode. Eight threads per process run were used in MPI/OpenMP hybrid calculations.

The speed-up of the whole HF calculations for the $TiO_2$ cluster is displayed in Figure 2. The calculations are performed using from 16 to 2048 CPU cores. The speed-up of 16 CPU cores is set to be 16, while the speed-up of 32 CPU cores is set to be 32 only for the original GAMESS of insulin. The speed-up of this work (only MPI) is improved in comparison with that of the original GAMESS (only MPI) because the initial guess calculation is simplified. Moreover, the speed-up of this work (MPI/OpenMP) is better than that of this work (only MPI) because the load balancing is improved by the MPI/OpenMP hybrid parallelization. It is surprising that the MPI/OpenMP calculation keeps the scaling performance even for 2048 CPU cores, though the parallel efficiency of the original GAMESS drops for 512 or 1024 CPU cores.

The MPI/OpenMP algorithm improves not only the speed-up but also the total computational (elapsed) time as shown in Table 1. The difference between the original GAMESS and this work (MPI/OpenMP) becomes large as the number of CPU cores increases. The MPI/OpenMP algorithm accelerates 1.6−3.7 times for 2048 CPU cores, while the computational times of the $TiO_2$ cluster for 16 CPU cores are almost the same. The MPI/OpenMP hybrid parallelization is achieved without increasing the computational cost.

Table 2 shows the computational time and speed-up of the Fock matrix generation for $TiO_2$ cluster and insulin. The speed-up of this work (MPI/OpenMP) is better than that of the original GAMESS, especially for over 512 CPU cores. For instance, the computational time of $TiO_2$ cluster is 174.8 s for 2048 CPU cores. Since the number of the SCF cycles is 30, the time per cycle is 5.8 s. This indicates that to make distributed tasks more equal by the hybrid parallelization is quite important to achieve extremely good speed-up.

The original GAMESS takes much more time than either of our two algorithms (only MPI, and MPI/OpenMP) for insulin.

***Table 1.*** Computational Time (Second) and Speed-up (in Parentheses) of Whole Hartree-Fock Calculation

| number of CPU cores | | 16 | 32 | 256 | 512 | 1024 | 2048 |
|---|---|---|---|---|---|---|---|
| | | | $TiO_2$ Cluster | | | | |
| original | only MPI | 18 176.4 | 9223.1 | 1368.6 | 806.8 | 527.6 | 383.5 |
| | | (16.0) | (31.5) | (212.5) | (360.5) | (551.2) | (758.3) |
| this work | only MPI | 18 045.6 | 9111.8 | 1241.2 | 695.2 | 428.7 | 273.7 |
| | | (16.0) | (31.7) | (232.6) | (415.3) | (673.5) | (1054.9) |
| this work | MPI/OpenMP | 18 121.6 | 9052.4 | 1214.6 | 656.5 | 381.1 | 234.2 |
| | | (16.0) | (32.0) | (238.7) | (441.7) | (760.8) | (1238.0) |
| | | | Insulin | | | | |
| original | only MPI | | 18 289.3 | 3629.9 | 2586.8 | 2054.1 | 1793.7 |
| | | | (32.0) | (161.2) | (226.2) | (284.9) | (326.3) |
| this work | only MPI | 21 988.8 | 11 157.5 | 1765.0 | 1073.4 | 721.7 | 540.9 |
| | | (16.0) | (31.5) | (199.3) | (327.8) | (487.5) | (650.4) |
| this work | MPI/OpenMP | 22 377.1 | 11 337.6 | 1654.1 | 975.6 | 642.1 | 480.4 |
| | | (16.0) | (31.6) | (216.5) | (367.0) | (557.6) | (745.3) |

***Table 2.*** Computational Time (Second) and Speed-up (in Parentheses) of Fock Matrix Generation

| number of CPU cores | | 16 | 32 | 256 | 512 | 1024 | 2048 |
|---|---|---|---|---|---|---|---|
| | | | $TiO_2$ Cluster | | | | |
| original | only MPI | 17 881.8 | 8984.9 | 1175.2 | 614.0 | 334.0 | 188.6 |
| | | (16.0) | (31.8) | (243.5) | (466.0) | (856.6) | (1517.0) |
| this work | only MPI | 17 953.5 | 9038.3 | 1175.2 | 627.9 | 360.0 | 203.1 |
| | | (16.0) | (31.8) | (244.4) | (457.5) | (797.9) | (1414.4) |
| this work | MPI/OpenMP | 17 777.6 | 8903.9 | 1150.4 | 597.2 | 316.4 | 174.8 |
| | | (16.0) | (31.9) | (247.3) | (476.3) | (899.0) | (1627.2) |
| | | | Insulin | | | | |
| original | only MPI | | 16 856.3 | 2491.9 | 1455.6 | 928.3 | 665.9 |
| | | | (32.0) | (216.5) | (370.6) | (581.1) | (810.0) |
| this work | only MPI | 21 454.4 | 10 827.8 | 1504.6 | 814.0 | 456.7 | 272.7 |
| | | (16.0) | (31.7) | (228.1) | (421.7) | (751.6) | (1258.8) |
| this work | MPI/OpenMP | 20 664.9 | 10 392.8 | 1384.3 | 734.8 | 410.7 | 253.9 |
| | | (16.0) | (31.8) | (238.8) | (450.0) | (805.1) | (1302.2) |

***Table 3.*** Computational Time (Second) and Ratio (in Parentheses) of Initial Guess Calculation

| number of CPU cores | | 16 | 32 | 256 | 512 | 1024 | 2048 |
|---|---|---|---|---|---|---|---|
| | | | $TiO_2$ Cluster | | | | |
| original | only MPI | 163.7 | 151.6 | 142.0 | 141.2 | 141.7 | 141.9 |
| | | (0.9%) | (1.6%) | (10.4%) | (17.5%) | (26.9%) | (37.0%) |
| this work | only MPI | 17.9 | 16.4 | 16.9 | 17.2 | 17.4 | 17.7 |
| | | (0.1%) | (0.2%) | (1.4%) | (2.5%) | (4.1%) | (6.5%) |
| this work | MPI/OpenMP | 16.5 | 13.7 | 11.9 | 12.0 | 12.2 | 12.3 |
| | | (0.1%) | (0.2%) | (1.0%) | (1.8%) | (3.2%) | (5.3%) |
| | | | Insulin | | | | |
| original | only MPI | | 695.6 | 667.8 | 671.6 | 674.0 | 667.8 |
| | | | (3.8%) | (18.4%) | (26.0%) | (32.8%) | (37.2%) |
| this work | only MPI | 74.8 | 72.8 | 73.2 | 73.7 | 74.2 | 74.6 |
| | | (0.3%) | (0.7%) | (4.1%) | (6.9%) | (10.3%) | (13.8%) |
| this work | MPI/OpenMP | 68.6 | 59.1 | 52.8 | 52.6 | 53.0 | 53.3 |
| | | (0.3%) | (0.5%) | (3.2%) | (5.4%) | (8.3%) | (11.1%) |

In the original, the if clause and the counter are used for the distribution of the 2-ERI calculation in the third loop, and all processes run until the if clause. A lot of 2-ERI are skipped by the Schwartz screening because of the tight basis set, STO-3G, and the computational cost for each integral is small. The cost for the if clause and the counter becomes relatively high, although that of the $TiO_2$ cluster is negligible because the computational cost of 2-ERI including *d* functions is large. Therefore, the developed algorithm can drastically reduce the computational time due to the simple distribution.

Table 3 summarizes the computational time and the ratio of the initial guess calculation for the $TiO_2$ cluster and the insulin. In the original, the initial guess calculation occupies about 37% of the total calculation time for 2048 CPU cores, though the ratio is less than 1% for 16 CPU cores. The ratio of this work (MPI/OpenMP) becomes about 5−11% even for 2048 CPU cores because of the simple orbital projection. The acceleration of the initial guess calculation contributes to the speed-up of the total time. The parallelization and acceleration of all steps are significant, and the initial guess calculation is not a bottleneck of parallel HF calculations for hundreds or thousands of CPU cores.

The computational time and the ratio of solving the Hartree−Fock equation are shown in Table 4, in which the diagonalization of the Fock matrix in the first and last iterations, the approximate SOSCF calculation in other

***Table 4.*** Computational Time (Second) and Ratio (in Parentheses) of Solving the Hartree-Fock Equation

| number of CPU cores | | 16 | 32 | 256 | 512 | 1024 | 2048 |
|---|---|---|---|---|---|---|---|
| | | | | $TiO_2$ Cluster | | | |
| original | only MPI | 128.4<br>(0.7%) | 84.4<br>(0.9%) | 49.8<br>(3.6%) | 49.6<br>(6.1%) | 50.0<br>(9.5%) | 51.1<br>(13.3%) |
| this work | only MPI | 71.9<br>(0.4%) | 55.2<br>(0.6%) | 47.4<br>(3.8%) | 48.3<br>(6.9%) | 49.8<br>(11.6%) | 51.4<br>(18.8%) |
| this work | MPI/OpenMP | 325.4<br>(1.8%) | 133.1<br>(1.5%) | 51.0<br>(4.2%) | 46.0<br>(7.0%) | 51.1<br>(13.4%) | 45.6<br>(19.5%) |
| | | | | Insulin | | | |
| original | only MPI | | 732.3<br>(4.0%) | 466.5<br>(12.9%) | 456.4<br>(17.6%) | 448.6<br>(21.8%) | 456.6<br>(25.5%) |
| this work | only MPI | 444.8<br>(2.0%) | 246.4<br>(2.2%) | 182.4<br>(10.3%) | 181.6<br>(16.9%) | 187.6<br>(26.0%) | 191.1<br>(35.3%) |
| this work | MPI/OpenMP | 1630.1<br>(7.3%) | 875.7<br>(7.7%) | 213.6<br>(12.9%) | 185.4<br>(19.0%) | 176.2<br>(27.4%) | 171.1<br>(35.6%) |

iterations, and the new density matrix generation are performed. The computational times of all algorithms are reduced as the number of CPU cores increases because the SOSCF calculation is originally parallelized by MPI. The difference of the computational time for insulin cluster is large for 2048 CPU cores. The dimension of the Fock and MO matrices is 1.5 times larger than that for the $TiO_2$ cluster, and cache misses easily occur in the density matrix generation because BLAS routines are not used in the original. The difference between only MPI and MPI/OpenMP of insulin is 20 s for 2048 CPU cores. This comes from the intranode parallelization of the Fock matrix diagonalization. This indicates that introducing internode parallelization of diagonalization is significant to reduce the computational time more.

## 4. Conclusion

We developed the new MPI/OpenMP hybrid parallel algorithm of the HF calculation and implemented it into the GAMESS program. The computational time and the speedup of the whole Hartree−Fock (HF) calculation are improved by introducing the MPI/OpenMP hybrid parallelization and the simple formula for the initial guess calculation. The basis sets used here are 6-31G and STO-3G. When a larger basis set is used, the hybrid parallelization effect is expected to be more important. The ratio of the initial guess calculation will decrease because the numbers of occupied orbitals and basis functions for the extended Hückel (EH) calculation are constant. The hybrid parallelization can reduce the amount of the memory use per node at the replicated data approach because all large matrices except for the Fock matrix are shared with all threads in a process.

For the Fock matrix generation, the better load balancing is obtained by the OpenMP dynamic distribution and the less MPI processes compared to that of the conventional MPI parallelization. The reduction of the computational time is also achieved by the distribution without the if clause. The ratio of the initial guess calculation becomes about 37% for 2048 CPU cores using the original GAMESS. The computational time and the ratio are drastically reduced by the use of the simple formula for the orbital projection and the parallelization of all steps. It is significant to apply basic linear algebra subprograms (BLAS) and linear algebra package (LAPACK) libraries of matrix multiplications and diagonalization for intranode parallelization and reduction of cache misses. The introduction of internode parallelization of diagonalization is necessary to accelerate more.

HF calculations of nanosize molecules without errors caused by modeling or approximations are now practical with hundreds or thousands of CPU cores. On the basis of the parallelization technique, algorithms for DFT and electron correlation calculations will be developed with high parallel efficiencies and large memory arrays, and most calculations based on the QM method will be accelerated.

## References

(1) Warshel, A.; Karplus, M. *J. Am. Chem. Soc.* **1972**, *94*, 5612–5625.

(2) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227–249.

(3) Maseras, F.; Morokuma, K. *J. Comput. Chem.* **1995**, *16*, 1170–1179.

(4) Dapprich, S.; Komaromi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J. *J. Mol. Struct. (THEOCHEM)* **1999**, *461*, 1–21.

(5) Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. *Chem. Phys. Lett.* **1999**, *313*, 701–706.

(6) Fedorov, D. G.; Kitaura, K. *J. Chem. Phys.* **2004**, *120*, 6832–6840.

(7) Chen, Z.; Nagase, S.; Hirsch, A.; Haddon, R. C.; Thiel, W.; Schleyer, P. v. R. *Angew. Chem., Int. Ed. Engl.* **2004**, *43*, 1552–1554.

(8) Fedorov, D. G.; Kitaura, K. *Chem. Phys. Lett.* **2004**, *389*, 129–134.

(9) Anderson, A. G.; Goddard, W. A., III; Schroder, P. *Comput. Phys. Commun.* **2007**, *177*, 298–306.

(10) Yasuda, K. *J. Chem. Theory Comput.* **2008**, *4*, 1230–1236.

(11) Yasuda, K. *J. Comput. Chem.* **2008**, *29*, 334–342.

(12) Vogt, L.; Olivares-Amaya, R.; Kermes, S.; Shao, Y.; Amador-Bedolla, C.; Aspuru-Guzik, A. *J. Phys. Chem. A* **2008**, *112*, 2049–2057.

(13) Ufimtsev, I. S.; Martinez, T. J. *J. Chem. Theory Comput.* **2008**, *4*, 222–231.

(14) Ufimtsev, I. S.; Martinez, T. J. *J. Chem. Theory Comput.* **2009**, *5*, 1004–1015.

(15) Brown, P.; Woods, C.; McIntosh-Smith, S.; Manby, F. R. *J. Chem. Theory Comput.* **2008**, *4*, 1620–1626.

(16) Whitten, J. L. *J. Chem. Phys.* **1973**, *58*, 4496–4501.

(17) TOP500 Supercomputing Sites; http://www.top500.org (accessed Mar 19, 2010).

(18) Nieplocha, J.; Harrison, R. J.; Littlefield, R. J. *Proceedings of the 1994 ACM/IEEE conference on Supercomputing,* **1994**, pp 340–349.

(19) Nieplocha, J.; Harrison, R. J.; Littlefield, R. J. *J. Supercomput.* **1996**, *10*, 197–220.

(20) Fletcher, G. D.; Schmidt, M. W.; Bode, B. M.; Gordon, M. S. *Comput. Phys. Commun.* **2000**, *128*, 190–200.

(21) Sosa, C. P.; Scalmani, G.; Gomperts, R.; Frisch, M. J. *Parallel Comput.* **2000**, *26*, 843–856.

(22) Woods, C. J.; Brown, P.; Manby, F. R. *J. Chem. Theory Comput.* **2009**, *5*, 1776–1784.

(23) Bolding, B.; Baldridge, K. *Comput. Phys. Commun.* **2000**, *128*, 55–66.

(24) Smith, L.; Kent, P. *Concurrency: Pract. Exper.* **2000**, *12*, 1121–1129.

(25) Shellman, S. D.; Lewis, J. P.; Glaesemann, K. R.; Sikorski, K.; Voth, G. A. *J. Comput. Phys.* **2003**, *188*, 1–15.

(26) Medvedev, D. M.; Goldfield, E. M.; Gray, S. K. *Comput. Phys. Commun.* **2005**, *166*, 94–108.

(27) Hutter, J.; Curioni, A. *Parallel Comput.* **2005**, *31*, 1–17.

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

(29) Nielsen, I. M. B.; Janssen, C. L. *Comput. Phys. Commun.* **2000**, *128*, 238–244.

(30) Janssen, C. L.; Nielsen, I. M. B.; Leininger, M. L.; Valeev, E. F.; Kenny, J. P.; Seidl, E. T. *The Massively Parallel Quantum Chemistry Program (MPQC)*; Sandia National Laboratories: Livemore, CA, 2008.

(31) Alexeev, Y.; Kendall, R. A.; Gordon, M. S. *Comput. Phys. Commun.* **2002**, *143*, 69–82.

(32) Takashima, H.; Yamada, S.; Obara, S.; Kitamura, K.; Inabata, S.; Miyakawa, N.; Tanabe, K.; Nagashima, U. *J. Comput. Chem.* **2002**, *23*, 1337–1346.

(33) Cremer, D.; Gauss, J. *J. Comput. Chem.* **1986**, *7*, 274–282.

(34) Baker, J.; Wolinski, K.; Malagoli, M.; Kinghorn, D.; Wolinski, P.; Magyarfalvi, G.; Saebo, S.; Janowski, T.; Pulay, P. *J. Comput. Chem.* **2009**, *30*, 317–335.

(35) Fischer, T. H.; Almlof, J. *J. Phys. Chem.* **1992**, *96*, 9768–9774.

(36) Chaban, G.; Schmidt, M. W.; Gordon, M. S. *Theor. Chem. Acc.* **1997**, *97*, 88–95.

(37) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347–1363.

(38) Hua, Q. X.; Shoelson, S. E.; Kochoyan, M.; Weiss, M. A. *Nature* **1991**, *354*, 238–241.

(39) Cuppen, J. J. M. *Numer. Math.* **1981**, *36*, 177–195.

CT100083W

JCTC Journal of Chemical Theory and Computation

# van der Waals Interactions in Density-Functional Theory: Intermolecular Complexes

Felix O. Kannemann and Axel D. Becke*

*Department of Chemistry, Dalhousie University, Halifax, Nova Scotia, Canada B3H 4J3*

Received December 30, 2009

**Abstract:** In previous work (*J. Chem. Theory Comput.* **2009**, *5*, 719), we assessed the performance of standard semilocal exchange-correlation density functionals plus the nonempirical dispersion model of Becke and Johnson (*J. Chem. Phys.* **2007**, *127*, 154108) on binding energy curves of rare-gas diatomics. The results were encouraging. In this work, we extend our study to 65 intermolecular complexes representing a wide variety of van der Waals interactions including dispersion, hydrogen bonding, electrostatic, and stacking. Comparisons are made with other density-functional methods for van der Waals interactions in the literature.

## 1. Introduction

An accurate description of van der Waals (vdW) interactions is required for electronic structure calculations on biomolecules, intermolecular complexes, molecular crystals, and polymers. Although conventional semilocal density-functional theory (DFT) gives accurate predictions for many molecular and solid-state properties, semilocal functionals are inherently unable to describe dispersion, a nonlocal correlation effect.[1] Thus, generalized gradient approximation (GGA), meta-GGA, and hybrid functionals are unreliable for systems where van der Waals interactions are important, even though they may give accidentally good results in limited cases.

Many methods have been developed to treat van der Waals interactions with DFT. These range from physically rigorous dispersion functionals derived from first principles to entirely empirical corrections or parametrizations. A comprehensive review of such methods is given by Johnson et al.[2] We provide an overview of the established methods and mention more recent approaches not covered in the review of Johnson et al.

The most rigorous description of dispersion interactions is provided by explicitly nonlocal correlation functionals. However, these methods are more computationally demanding and complicated than standard DFT. An example is the Andersson−Langreth−Lundqvist (ALL) functional[3] for nonoverlapping systems, also derived by Dobson and Dinte (DD),[4] which has been applied to intermolecular complexes in conjunction with long-range corrected (LC) exchange-correlation functionals.[5,6] A seamless van der Waals density functional (vdW-DF), valid at all interatomic distances, has been developed by Langreth and co-workers[7,8] and applied to many molecular, solid-state, and biochemical systems.[9] Other approaches use *ab initio* methods such as MP2 or CCSD(T) to describe long-range electron correlation, which is combined with short-range DFT.[10−13] The advantages of these range-separated hybrid (RSH) methods compared to standard MP2 or CCSD(T) are reduced basis-set dependence and smaller basis-set superposition error (BSSE).

An entirely empirical approach to dispersion interactions involves the parametrization of highly flexible density functionals by including van der Waals complexes in their fitting sets, in spite of the failure of semilocal DFT to capture dispersion physics. Early functionals of this type such as X3LYP,[14] M05, and M05−2X[15] give large errors for stacking interactions,[16−18] while the newer M06−2X functional performs better due to additional empirical parameters.[19] Also, these functionals do not improve the description of prototypical vdW systems such as rare-gas diatomics, compared to the standard GGAs B97−1 and PBE.[20] Neither are they able to reproduce the asymptotic $R^{-6}$ behavior of the dispersion energy at large separation.[2,17]

Theoretically more sound, but still highly empirical, is the addition of explicit dispersion corrections to conventional functionals. These can take the form of $C_6R^{-6}$ corrections (DFT-D)[21−27] or atom-centered dispersion-correcting pseudopotentials (DCP).[28−31] Both approaches are easy to implement in existing electronic structure codes and have negli-

* Corresponding author e-mail: axel.becke@dal.ca.

10.1021/ct900699r © 2010 American Chemical Society
Published on Web 03/26/2010

gible computational cost. However, the necessary $C_6$ coefficients and vdW radii (DFT-D), or pseudopotential parameters (DCP), must be empirically determined for each element and have limited transferability. Real dispersion coefficients depend on the oxidation state of an atom and its molecular environment, which is disregarded in the DFT-D and DCP approaches.[2,32]

Much more satisfying dispersion corrections are now available using system-dependent dispersion coefficients calculated from first principles. Three recently proposed methods provide dispersion coefficients from the ground-state electron density of the system. Tkatchenko and Scheffler[32] use Hirshfeld atomic volumes to calculate atom-in-molecule dispersion coefficients from corresponding free-atom dispersion coefficients. Their method yields accurate $C_6$ coefficients but relies on free-atom reference data for dispersion coefficients and polarizabilities. Sato and Nakai[33] have formulated the local response dispersion (LRD) method, which evaluates polarizabilites and dispersion coefficients from first principles without the need for free-atom reference data. It uses the same local density approximation to the response function as in the DD/ALL and vdW-DF approaches. A multipole expansion of the dispersion energy is introduced, and numerical evaluation of a Casimir−Polder imaginary frequency integral is used to obtain dispersion coefficients,[33] replacing the double numerical integrations required in the ALL and vdW-DF methods.

Preceeding both the Tkatchenko−Scheffler and LRD methods is the nonempirical dispersion model of Becke and Johnson.[34] This model generates dispersion coefficients from the exchange-hole dipole moment (XDM)[35] using occupied orbitals or the electron density.[36] Its theoretical foundations have been investigated by various other authors.[37−39] In addition to the exchange-hole dipole moment, the method uses atom-in-molecule polarizabilites derived from atomic reference data and Hirshfeld atomic volumes.[34] A modification introduced by Krishtal et al.[40] employs intrinsic atomic polarizabilites, obtained by a Hirshfeld partitioning of molecular polarizibility tensors, which also gives anisotropy corrections. Kong et al.[41] recently implemented the XDM model self-consistently and assessed the importance of self-consistency on the calculation of dispersion energies and forces.

The XDM dispersion model is part of the DF07 functional,[42] a universal DFT for thermochemistry, kinetics, and van der Waals interactions.[43] DF07 is exact-exchange-based, using 100% Hartree−Fock (HF) exchange. HF exchange is computationally expensive, however, and a combination of XDM dispersion with a pure GGA functional is desirable. Unfortunately, standard exchange GGAs give everything from artificial binding (e.g., LDA, PW91,[44,45] PBE,[46] and B86[47]) to strong over-repulsion (B88[48]) in rare-gas diatomic tests, compared to HF repulsion.[49]

In a previous paper,[50] we benchmarked standard exchange GGAs for their ability to reproduce HF repulsion in 10 rare-gas diatomics and found that the nonempirical exchange GGA of Perdew and Wang (PW86)[51,52] performs best. This conclusion was reached by comparing GGA exchange-only interaction energies, at equilibrium separations, to exact Hartree−Fock interaction energies. More recently, Murray, Lee, and Langreth[53] have examined standard exchange GGAs in interacting molecular systems (dimers of $H_2$, $N_2$, $CO_2$, ammonia, methane, ethene, benzene, and pyrazine). They also find that PW86 best reproduces Hartree−Fock repulsion energies over a range of intermolecular separations.

We then combined[50] the PW86 exchange functional with the PBE[46] correlation functional and the XDM dispersion model as follows,

$$E_{\mathrm{XC}} = E_X^{PW86} + E_C^{\mathrm{PBE}} + E_{\mathrm{disp}}^{\mathrm{XDM}} \tag{1}$$

and obtained excellent binding energy curves for our 10 rare-gas diatomics. In the present work, we extend the benchmarking of eq 1 from rare-gas diatomics to intermolecular complexes. A comprehensive test set of 65 complexes has been assembled (see section 2), containing vdW interactions from $He_2$ through electrostatic, hydrogen bonding, and stacking interactions of importance in biochemistry. With only two parameters, an excellent fit is obtained to binding energies spanning 3 orders of magnitude in strength.

In section 3, we compare our results with results from a variety of other methods in the literature. Our method compares quite favorably, especially considering its small number (2) of fitted parameters.

## 2. Fitting of Dispersion Damping Parameters

In the XDM model of Becke and Johnson, the dispersion energy is given by

$$E_{\mathrm{disp}}^{\mathrm{XDM}} = -\frac{1}{2}\sum_{i\neq j}\left(\frac{C_{6,ij}}{R_{\mathrm{vdW},ij}^6 + R_{ij}^6} + \frac{C_{8,ij}}{R_{\mathrm{vdW},ij}^8 + R_{ij}^8} + \frac{C_{10,ij}}{R_{\mathrm{vdW},ij}^{10} + R_{ij}^{10}}\right) \tag{2}$$

The nonempirical, system-dependent dispersion coefficients $C_{6,ij}$, $C_{8,ij}$, and $C_{10,ij}$ are obtained from the exchange-hole dipole moment and atom-in-molecule polarizabilities using second-order perturbation theory.[35] In the "exact-exchange" (XX) version of the XDM model, the dipole moment of the exchange hole is calculated using occupied orbitals.[34] Alternatively, the Becke−Roussel density-functional model of the exchange hole[54] can be used to calculate an approximate dipole moment, giving the "BR" variant of the XDM model.[36]

The van der Waals separations $R_{\mathrm{vdW},ij}$ in eq 2 are assumed to be linearly related to "critical" interatomic separations $R_{\mathrm{c},ij}$ by

$$R_{\mathrm{vdW},ij} = a_1 R_{\mathrm{c},ij} + a_2 \tag{3}$$

where $a_1$ and $a_2$ are universal parameters and $R_{\mathrm{c},ij}$ is the average value of the ratios $(C_{8,ij}/C_{6,ij})^{1/2}$, $(C_{10,ij}/C_{6,ij})^{1/4}$, and $(C_{10,ij}/C_{8,ij})^{1/2}$. At this separation, the three asymptotic dispersion terms are approximately equal to each other:

$$\frac{C_{6,ij}}{R_{\mathrm{c},ij}^6} \approx \frac{C_{8,ij}}{R_{\mathrm{c},ij}^8} \approx \frac{C_{10,ij}}{R_{\mathrm{c},ij}^{10}} \tag{4}$$

and the asymptotic expansion of the dispersion energy is no longer valid.[34]

In our previous work,[50] the $a_1$ and $a_2$ parameters in eq 3 were fit to the binding energies of 10 rare-gas diatomics. Good performance in rare-gas systems does not, however, guarantee good performance in intermolecular complexes.[16,55] In this work, we therefore fit the damping parameters to a larger set of 65 complexes. This set includes the following:

- The 22 complexes of the "S22" biochemical benchmark set.[55] S22 uses CCSD(T) or MP2 geometries, and the binding energies are CCSD(T)/complete basis-set estimates. Monomer deformations are not considered.
- Ten rare-gas diatomics involving He through Kr. We use the experimentally derived data of Tang and Toennies (TT).[56]
- Twelve complexes from the NC31/05 "non-covalent" database of Zhao and Truhlar,[20,57] excluding those systems duplicated in the S22 and TT sets. The NC31/05 database uses mainly MC-QCISD/3 geometries and W1 binding energies, including monomer deformation energies. We also exclude charge-transfer complexes from our training set, as GGAs strongly overestimate charge-transfer interactions due to severe self-interaction error.[58,59] This error is partly removed by hybrid functionals that mix in a fraction of HF exchange,[58,59] and more completely by LC-hybrid methods that use 100% long-range HF exchange.[27,60] However, the focus of the present work is on a pure GGA functional without inclusion of HF exchange.
- Twenty-one systems from the 45 vdW complexes of Johnson and Becke (JB),[34] excluding systems contained in the preceding databases. Binding energies for the JB systems are mainly at the estimated CCSD(T)/complete basis-set limit and do not include monomer deformation energies.

This compilation of reference data from various sources comprises a diverse set of intermolecular complexes with binding energies ranging from 0.022 kcal/mol ($He_2$) to 20.65 kcal/mol (hydrogen bonded uracil dimer) and including dispersion, hydrogen bonding, electrostatic, and stacking interactions.

The damping parameters $a_1$ and $a_2$ are determined by minimizing the root-mean-square percent error (RMS%E)

$$\mathrm{RMS\%E} = 100 \times \sqrt{\frac{1}{N}\sum_i^N\left(\frac{\mathrm{BE}_i^{\mathrm{calc}} - \mathrm{BE}_i^{\mathrm{ref}}}{\mathrm{BE}_i^{\mathrm{ref}}}\right)^2}$$

of our calculated binding energies $\mathrm{BE}_i^{\mathrm{calc}}$ with respect to the reference binding energies $\mathrm{BE}_i^{\mathrm{ref}}$ at the reference geometries. Cartesian coordinates for the complexes and monomers of this training set are provided in the Supporting Information. Our calculations were performed with the fully numerical, basis-set-free Numol program of Becke and Dickson[61-63] using LDA orbitals (i.e., "post-LDA") and the Perdew-Wang uniform-electron-gas exchange-correlation parametrization.[64] We use numerical grids of 302 angular points per atom and 80 ($Z \leq 2$), 120 ($2 < Z \leq 10$), 160 ($10 < Z \leq 18$), and 200 ($18 < Z \leq 36$) radial shells per atom.

The binding energies of the 65 complexes are shown in Table 1. Binding energies are taken to be positive quantities, i.e., negative values indicate a repulsive interaction. Table 1 also lists the dispersion contribution to the binding energy, calculated as

$$\%\mathrm{disp} = 100 \times \frac{\mathrm{BE(PW86PBE\text{-}XDM)} - \mathrm{BE(PW86PBE)}}{\mathrm{BE(PW86PBE\text{-}XDM)}}$$

Hydrogen bonded complexes have dispersion contributions < 20% and dipolar and "mixed" interactions < 75% (with the exception of the T-shaped benzene dimer), while in dispersion-bound and "stacked" complexes, the contribution of the dispersion energy exceeds 50%.

Table 2 contains the optimized $a_1$ and $a_2$ values and error statistics. As can be seen from Tables 1 and 2, PW86PBE describes hydrogen bonding and dipolar (electrostatic) interactions well but fails for dispersion. The addition of the XDM dispersion energy gives accurate binding energies for the whole set of 65 complexes. A few systems ($C_2H_4 \cdot HF$, $HF \cdot HF$, $NH_3 \cdot H_2O$, $H_2S \cdot HCl$, and $CH_3SH \cdot HCl$) are slightly overbound by PW86PBE itself, and addition of the dispersion energy worsens the agreement with the reference binding energies.

Table 2 also shows that, for this set of 65 complexes, the BR variant of the XDM dispersion model is significantly more accurate than the XX version. The opposite result was found in our work on rare-gas diatomics.[50] This can be understood by considering how well the exchange hole in the XDM dispersion model actually approximates the full exchange-correlation (XC) hole.[37,38] In rare-gas systems, which do not have nondynamical correlation, the exact-exchange hole (XX) is apparently a better approximation of the XC hole than the approximate BR hole. In molecular systems, however, nondynamical (left-right) correlation leads to a multicenter-to-single-center localization of the XC hole.[65,66] As the localized XC holes in molecules are more effectively modeled by semilocal (meta-)GGAs such as BR than by the delocalized exact-exchange hole, the BR version of the XDM dispersion model can be expected to work better in intermolecular complexes. Dynamical correlation also contributes to the XC hole, but the dipole moment of the XC hole should be rather insensitive to the effects of dynamical correlation,[38] thus justifying the use of the exchange-only hole in the XDM dispersion model.

Table 3 contains the $a_1$ and $a_2$ damping parameters for the PW86PBE-XDM functional obtained in our previous work[50] on rare-gas diatomics. Using these rare-gas-optimized parameters to calculate binding energies for the current set of 65 vdW complexes, we obtain similar error statistics (Table 3) compared to the fit in Table 2. In other words, the damping parameters optimized for rare-gas systems are transferable to more complex intermolecular interactions. Conversely, the damping parameters obtained in this work give good results for the binding energies of rare-gas diatomics, with mean absolute percentage errors (MA%E) of 24.7% for XDM(XX) and 10.3% for XDM(BR). This is very gratifying. The functional of eq 1, with the damped XDM dispersion model of eq 2, is apparently universally applicable to vdW interactions spanning 3 orders of magnitude in strength, with only two fitted parameters.

## 3. Performance on the S22 Benchmark Set

The "S22" database of Jurecka et al.[55] contains 22 intermolecular complexes of biochemical interest and covers hy-

***Table 1.*** Binding Energies (kcal/mol), Binding Energy Errors (kcal/mol), and Dispersion Contribution to Binding Energy (%) for the Training Set of 65 vdW Complexes

| | | | | PW86PBE | | PW86PBE-XDM(XX) | | | PW86PBE-XDM(BR) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| complex | database | type | $BE^{ref}$ | $BE^{calc}$ | error | $BE^{calc}$ | error | % disp | $BE^{calc}$ | error | % disp |
| He·He | TT | dispersion | 0.022 | −0.016 | −0.038 | 0.018 | −0.004 | 189 | 0.018 | −0.004 | 189 |
| He·Ne | TT | dispersion | 0.041 | −0.012 | −0.053 | 0.046 | 0.005 | 126 | 0.048 | 0.007 | 125 |
| He·Ar | TT | dispersion | 0.059 | −0.023 | −0.082 | 0.070 | 0.011 | 133 | 0.059 | 0.000 | 139 |
| He·Kr | TT | dispersion | 0.062 | −0.020 | −0.082 | 0.082 | 0.020 | 124 | 0.065 | 0.003 | 131 |
| Ne·Ne | TT | dispersion | 0.084 | −0.007 | −0.091 | 0.094 | 0.010 | 107 | 0.100 | 0.016 | 107 |
| Ne·Ar | TT | dispersion | 0.132 | −0.016 | −0.148 | 0.165 | 0.033 | 110 | 0.143 | 0.011 | 111 |
| Ne·Kr | TT | dispersion | 0.141 | −0.008 | −0.149 | 0.199 | 0.058 | 104 | 0.164 | 0.023 | 105 |
| Ar·Ar | TT | dispersion | 0.285 | −0.063 | −0.348 | 0.345 | 0.060 | 118 | 0.255 | −0.030 | 125 |
| Ar·Kr | TT | dispersion | 0.333 | −0.065 | −0.398 | 0.435 | 0.102 | 115 | 0.311 | −0.022 | 121 |
| Kr·Kr | TT | dispersion | 0.400 | −0.071 | −0.471 | 0.552 | 0.152 | 113 | 0.381 | −0.019 | 119 |
| He·N2 L-shaped | JB | dispersion | 0.053 | −0.026 | −0.079 | 0.041 | −0.012 | 163 | 0.040 | −0.013 | 165 |
| He·N2 T-shaped | JB | dispersion | 0.066 | −0.033 | −0.099 | 0.064 | −0.002 | 152 | 0.062 | −0.004 | 153 |
| He·FCl | JB | dispersion | 0.097 | −0.031 | −0.128 | 0.074 | −0.023 | 142 | 0.077 | −0.020 | 140 |
| FCl·He | JB | dispersion | 0.182 | 0.016 | −0.166 | 0.187 | 0.005 | 91 | 0.157 | −0.025 | 90 |
| CH4·C2H4 | JB | dispersion | 0.50 | 0.11 | −0.39 | 0.63 | 0.13 | 83 | 0.67 | 0.17 | 84 |
| CF4·CF4 | JB | dispersion | 0.78 | −0.41 | −1.19 | 0.50 | −0.28 | 181 | 0.70 | −0.08 | 158 |
| SiH4·CH4 | JB | dispersion | 0.81 | −0.06 | −0.87 | 0.89 | 0.08 | 106 | 0.88 | 0.07 | 106 |
| CO2·CO2 | JB | dispersion | 1.37 | 0.28 | −1.09 | 1.19 | −0.18 | 76 | 1.15 | −0.22 | 76 |
| OCS·OCS | JB | dispersion | 1.40 | −0.10 | −1.50 | 1.69 | 0.29 | 106 | 1.38 | −0.02 | 107 |
| C10H8·C10H8 parallel | JB | dispersion (stacking) | 3.78 | −2.70 | −6.48 | 2.87 | −0.91 | 194 | 4.50 | 0.72 | 160 |
| C10H8·C10H8 parallel crossed | JB | dispersion (stacking) | 5.28 | −2.46 | −7.74 | 3.95 | −1.33 | 162 | 5.85 | 0.57 | 142 |
| C10H8·C10H8 T-shaped | JB | dispersion (stacking) | 4.34 | −0.39 | −4.73 | 3.38 | −0.96 | 112 | 4.46 | 0.12 | 109 |
| C10H8·C10H8 T-shaped crossed | JB | dispersion (stacking) | 3.09 | −0.21 | −3.30 | 2.62 | −0.47 | 108 | 3.50 | 0.41 | 106 |
| CH4·NH3 | JB | dipole−induced dipole | 0.73 | 0.54 | −0.19 | 0.95 | 0.22 | 43 | 0.97 | 0.24 | 45 |
| SiH4·HF | JB | dipole−induced dipole | 0.73 | 0.17 | −0.56 | 0.63 | −0.10 | 73 | 0.62 | −0.11 | 73 |
| CH4·HF | JB | dipole−induced dipole | 1.65 | 1.29 | −0.36 | 1.77 | 0.12 | 27 | 1.76 | 0.11 | 27 |
| C2H4·HF | JB | dipole−induced dipole | 4.47 | 4.50 | 0.03 | 5.16 | 0.69 | 13 | 5.16 | 0.69 | 13 |
| CH3F·CH3F | JB | dipole−dipole | 2.33 | 1.32 | −1.01 | 2.11 | −0.22 | 37 | 2.16 | −0.17 | 39 |
| H2CO·H2CO | JB | dipole−dipole | 3.37 | 1.95 | −1.42 | 3.03 | −0.34 | 36 | 2.99 | −0.38 | 35 |
| CH3CN·CH3CN | JB | dipole−dipole | 6.16 | 4.40 | −1.76 | 6.16 | 0.00 | 29 | 6.13 | −0.04 | 28 |
| HCN·HF | JB | hydrogen bonding | 7.3 | 7.27 | −0.03 | 7.75 | 0.45 | 6 | 7.71 | 0.41 | 6 |
| (NH3)2 [C2h] | S22 | hydrogen bonding | 3.17 | 2.60 | −0.57 | 3.28 | 0.11 | 21 | 3.24 | 0.07 | 20 |
| (H2O)2 [Cs] | S22 | hydrogen bonding | 5.02 | 4.73 | −0.29 | 5.26 | 0.24 | 10 | 5.23 | 0.21 | 9 |
| formic acid dimer [C2h] | S22 | hydrogen bonding | 18.61 | 17.38 | −1.23 | 19.08 | 0.47 | 9 | 19.13 | 0.52 | 9 |
| formamide dimer [C2h] | S22 | hydrogen bonding | 15.96 | 14.30 | −1.67 | 16.04 | 0.08 | 11 | 16.09 | 0.13 | 11 |
| uracil dimer [C2h] | S22 | hydrogen bonding | 20.65 | 17.98 | −2.68 | 19.91 | −0.74 | 10 | 20.28 | −0.37 | 11 |
| 2-pyridoxine·2-aminopyridine [C1] | S22 | hydrogen bonding | 16.71 | 14.68 | −2.03 | 16.97 | 0.26 | 13 | 17.44 | 0.72 | 16 |
| adenine·thymine WC [C1] | S22 | hydrogen bonding | 16.37 | 13.65 | −2.72 | 16.11 | −0.26 | 15 | 16.68 | 0.31 | 18 |
| (CH4)2 [D3d] | S22 | dispersion | 0.53 | −0.14 | −0.67 | 0.52 | −0.01 | 128 | 0.54 | 0.01 | 127 |
| (C2H4)2 [D2d] | S22 | dispersion | 1.51 | 0.01 | −1.50 | 1.41 | −0.10 | 99 | 1.36 | −0.15 | 99 |
| benzene·CH4 [C3] | S22 | dispersion | 1.50 | −0.18 | −1.68 | 1.25 | −0.25 | 114 | 1.48 | −0.02 | 112 |
| benzene dimer parallel displaced [C2h] | S22 | dispersion (stacking) | 2.73 | −2.22 | −4.95 | 1.85 | −0.88 | 220 | 2.63 | −0.10 | 184 |
| pyrazine dimer [Cs] | S22 | dispersion (stacking) | 4.42 | −1.02 | −5.44 | 3.13 | −1.29 | 133 | 3.69 | −0.73 | 128 |
| uracil dimer stack [C2] | S22 | dispersion (stacking) | 10.12 | 2.38 | −7.74 | 7.62 | −2.50 | 69 | 8.78 | −1.35 | 73 |
| indole·benzene stack [C1] | S22 | dispersion (stacking) | 5.22 | −2.55 | −7.77 | 3.02 | −2.20 | 184 | 4.27 | −0.95 | 160 |
| adenine·thymine stack [C1] | S22 | dispersion (stacking) | 12.23 | 0.95 | −11.28 | 8.30 | −3.93 | 89 | 10.07 | −2.16 | 91 |
| ethene·ethyne [C2v] | S22 | mixed | 1.53 | 0.97 | −0.57 | 1.72 | 0.19 | 44 | 1.71 | 0.18 | 43 |
| benzene·H2O [Cs] | S22 | mixed | 3.28 | 1.69 | −1.59 | 3.01 | −0.27 | 44 | 3.17 | −0.11 | 47 |
| benzene·NH3 [Cs] | S22 | mixed | 2.35 | 0.65 | −1.70 | 2.05 | −0.30 | 68 | 2.22 | −0.13 | 71 |

***Table 1.*** Continued

| | | | | PW86PBE | | PW86PBE-XDM(XX) | | | PW86PBE-XDM(BR) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| complex | database | type | $BE^{ref}$ | $BE^{calc}$ | error | $BE^{calc}$ | error | % disp | $BE^{calc}$ | error | % disp |
| benzene•HCN [Cs] | S22 | mixed | 4.46 | 2.35 | −2.12 | 4.01 | −0.45 | 42 | 4.15 | −0.31 | 43 |
| benzene dimer T-shaped [C2v] | S22 | mixed | 2.74 | −0.24 | −2.98 | 1.94 | −0.80 | 112 | 2.40 | −0.34 | 110 |
| indole•benzene T-shaped [C1] | S22 | mixed | 5.73 | 1.45 | −4.28 | 4.31 | −1.42 | 66 | 4.95 | −0.79 | 71 |
| phenol dimer [C1] | S22 | mixed | 7.05 | 3.62 | −3.43 | 6.00 | −1.05 | 40 | 6.52 | −0.53 | 45 |
| HF•HF | HB6/04 | hydrogen bonding | 4.57 | 4.59 | 0.02 | 4.92 | 0.35 | 7 | 4.91 | 0.34 | 7 |
| NH3•H2O | HB6/04 | hydrogen bonding | 6.41 | 6.52 | 0.11 | 7.14 | 0.73 | 9 | 7.12 | 0.71 | 8 |
| H2S•H2S | DI6/04 | dipole−dipole | 1.66 | 1.29 | −0.37 | 2.22 | 0.56 | 42 | 2.02 | 0.36 | 36 |
| HCl•HCl | DI6/04 | dipole−dipole | 2.01 | 1.81 | −0.21 | 2.66 | 0.65 | 32 | 2.43 | 0.42 | 26 |
| H2S•HCl | DI6/04 | dipole−dipole | 3.35 | 3.66 | 0.31 | 4.70 | 1.35 | 22 | 4.43 | 1.08 | 17 |
| CH3Cl•HCl | DI6/04 | dipole−dipole | 3.55 | 2.97 | −0.58 | 4.34 | 0.79 | 32 | 4.06 | 0.51 | 27 |
| HCN•CH3SH | DI6/04 | dipole−dipole | 3.59 | 3.11 | −0.48 | 4.33 | 0.74 | 28 | 4.20 | 0.61 | 26 |
| CH3SH•HCl | DI6/04 | dipole−dipole | 4.16 | 4.89 | 0.73 | 6.46 | 2.30 | 24 | 6.14 | 1.98 | 20 |
| CH4•Ne | WI7/05 | dispersion | 0.22 | −0.04 | −0.26 | 0.14 | −0.08 | 126 | 0.14 | −0.08 | 124 |
| C6H6•Ne | WI7/05 | dispersion | 0.47 | −0.14 | −0.61 | 0.27 | −0.20 | 152 | 0.35 | −0.13 | 140 |
| C2H2•C2H2 | PPS5/05 | dispersion (stacking) | 1.34 | 0.88 | −0.46 | 1.73 | 0.39 | 49 | 1.64 | 0.30 | 46 |
| C6H6•C6H6 parallel | PPS5/05 | dispersion (stacking) | 1.81 | −2.02 | −3.83 | 0.82 | −0.99 | 347 | 1.49 | −0.32 | 235 |

***Table 2.*** Optimized Dispersion Damping Parameters and Error Statistics for the Training Set of 65 vdW Complexes

| dispersion | none | XDM(XX) | XDM(BR) |
|---|---|---|---|
| $a_1$ | | 0.68 | 0.82 |
| $a_2$ (Å) | | 1.43 | 1.16 |
| RMS%E (%) | 96.8 | 24.0 | 15.8 |
| MA%E (%) | 79.1 | 19.9 | 12.6 |
| MAE (kcal/mol) | 1.72 | 0.53 | 0.33 |
| MaxE(−)<br>(kcal/mol) | −11.28<br>(A•T stack) | −3.93<br>(A•T stack) | −2.16<br>(A•T stack) |
| MaxE(+)<br>(kcal/mol) | 0.73<br>(CH3SH•HCl) | 2.30<br>(CH3SH•HCl) | 1.98<br>(CH3SH•HCl) |

***Table 3.*** Error Statistics for the Current Set of 65 vdW Complexes Using Rare-Gas-Optimized Damping Parameters[50]

| dispersion | XDM(XX) | XDM(BR) |
|---|---|---|
| $a_1$ | 0.95 | 0.75 |
| $a_2$(Å) | 0.87 | 1.25 |
| RMS%E (%) | 30.5 | 18.9 |
| MA%E (%) | 23.3 | 14.2 |
| MAE (kcal/mol) | 0.73 | 0.35 |
| MaxE(−) (kcal/mol) | −5.95 (A•T stack) | −0.76 (A•T stack) |
| MaxE(+) (kcal/mol) | 1.74 (CH3SH•HCl) | 2.26 (CH3SH•HCl) |

drogen bonding, dispersion, and stacking interactions. It provides CCSD(T) binding energies at the estimated complete basis-set limit and has been widely adopted to assess the performance of electronic structure methods for intermolecular interactions. In Table 4, we list mean absolute errors (MAE, kcal/mol) and mean absolute percent errors (MA%E) for a variety of DFT methods for which benchmark data on the S22 set are available in the literature.[17,19,27,30,32,33,67−71]

The Becke−Roussel variant of the XDM dispersion model, XDM(BR), gives excellent binding energies for the S22 set as demonstrated by its low MAE and MA%E values. Its accuracy is comparable to the empirical DFT-D methods, the highly parametrized M06-2X and $\omega$B97X functionals, and the much more expensive "double hybrid" functionals (which include nonlocal correlation through second order MP2 perturbation theory). By coincidence, our previous damping parameter fit to rare-gas systems[50] (denoted as "TT" in Table 4) gives slightly better error statistics than the current fit to 65 intermolecular complexes. We also note that the exact-exchange version of the XDM dispersion model, XDM(XX), is much less accurate for the S22 set for reasons explained in section 2. Given its higher accuracy and lower computational cost, we prefer the XDM(BR) variant over XDM(XX). XDM(BR) is also the method which was recently implemented self-consistently.[41]

The nonempirical dispersion approaches of Sato and Nakai (LRD) and Tkatchenko and Scheffler (TS) also give excellent binding energies for the S22 set, as do the empirical (DFT-D) dispersion corrections. The van der Waals density functional (vdW-DF) is less accurate, and as shown by Gulans et al.[67] and Klimes et al.,[68] the results depend on the underlying exchange functional. With revPBE[72] exchange, the complexes of the S22 set are systematically underbound,[67,68]

***Table 4.*** Mean Absolute Errors (MAE, kcal/mol) of Various DFT Methods for the S22 Set and the Subsets of Hydrogen-Bonded (HB), Dispersion-Dominated (disp), and Mixed (mix) Complexes, As Well As Mean Absolute Percentage Errors (MA%E) for the S22 Set[a]

| method | type | MAE (S22) | MAE (HB) | MAE (disp) | MAE (mix) | MA%E |
|---|---|---|---|---|---|---|
| PW86PBE-XDM(BR) | GGA-XDM | 0.46 | 0.33 | 0.68 | 0.34 | 7.4 |
| PW86PBE-XDM(BR) TT[50] | GGA-XDM | 0.31 | 0.52 | 0.27 | 0.14 | 6.2 |
| PW86PBE-XDM(XX) | GGA-XDM | 0.81 | 0.31 | 1.39 | 0.64 | 14.3 |
| PW86PBE-XDM(XX) TT[50] | GGA-XDM | 1.33 | 0.43 | 2.39 | 1.03 | 26.7 |
| *vdW-DF(revPBE)*[67] | GGA+vdW-DF | 1.39 | 2.81 | 0.79 | 0.65 | 18.3 |
| vdW-DF(B86)[68] | GGA+vdW-DF | 0.53 | 0.76 | 0.58 | 0.23 | na |
| LC-BOP+LRD(6 + 8+10)[33] | LC hybrid GGA+LRD | 0.27 | 0.35 | 0.20 | 0.28 | 5.7 |
| PBE+TS[32] | GGA+TS | 0.30 | 0.46 | 0.30 | 0.14 | na |
| ωB97X-D[27] | LC hybrid GGA-D | 0.22 | 0.24 | 0.26 | 0.17 | 5.4 |
| B97-D[27] | GGA-D | 0.50 | 0.84 | 0.43 | 0.24 | 6.4 |
| B3LYP-D[27] | hybrid GGA-D | 0.48 | 0.81 | 0.35 | 0.28 | 8.5 |
| BLYP-D[27] | GGA-D | 0.33 | 0.28 | 0.52 | 0.16 | 8.6 |
| revPBE+LAP[30] | GGA-DCP(LAP) | 0.57 | 1.11 | 0.42 | 0.22 | 7.0 |
| *B3LYP-DCP*[69] | hybrid GGA-DCP | 0.93 | 1.34 | 0.90 | 0.56 | 20.4 |
| M06−2X[19] | hybrid meta-GGA | 0.40 | 0.70 | 0.17 | 0.35 | 6.4 |
| M05−2X[17] | hybrid meta-GGA | 0.86 | 0.75 | 1.26 | 0.53 | 14.8 |
| mPW2PLYP-D[71] | double hybrid GGA-D | 0.46 | 0.50 | 0.70 | 0.16 | 8.4 |
| B2PLYP-D[71] | double hybrid GGA-D | 0.27 | 0.18 | 0.48 | 0.12 | 6.6 |
| ωB97X-2(LP)[70] | LC double hybrid GGA | 0.24 | 0.21 | 0.30 | 0.22 | 7.4 |

[a] Methods using optimized geometries instead of the S22 reference geometries are shown in *italics*.

while the combination of B86[47] exchange and vdW-DF gives substantially improved results.[68]

A similar case is B3LYP-DCP,[69] which combines the dispersion-correcting pseudopotentials of DiLabio and Mackie[73] with the B3LYP[74,75] functional. Due to the use of overly repulsive B88 exchange[48] in B3LYP, B3LYP-DCP underbinds all systems in the S22 database except for the water dimer.[69] The revPBE+LAP[30] method, which combines the revPBE[72] exchange-correlation functional with a dispersion-correcting local atomic potential (LAP),[30] does not suffer from this deficiency but is still rather inaccurate for hydrogen-bonded complexes.

A variety of DFT methods reproduce the binding energies of the intermolecular complexes in the S22 database very well. Those that use nonempirical dispersion coefficients and only a few parameters in their damping functions are the present XDM(BR), the LRD of Sato and Nakai, and the Tkatchenko−Scheffler methods. The nonempirical vdW-DF is comparably less accurate with the original revPBE exchange but can be improved by changing the underlying exchange functional.[68] The influence of the exchange functional is well-known[49,76,77] but has been somewhat overlooked until the recent studies of refs 50, 53, and 68 and the present work.

## 4. Conclusions

We have shown that the XDM dispersion model of Becke and Johnson can be combined with standard GGAs for exchange (PW86) and correlation (PBE) to give an excellent description of van der Waals interactions. The XDM dispersion model contains only two empirical parameters in the damping function. These have been fit to a set of 65 complexes ranging from rare-gas systems to nucleic acid base pairs and spanning 3 orders of magnitude in binding energy strength. Also, the dispersion damping parameters optimized for rare-gas diatomics in our previous work[50] are found to be highly transferable to the larger set of intermolecular interactions.

The Becke−Roussel variant of XDM, XDM(BR), is more accurate for intermolecular complexes than the exact-exchange-hole variant, XDM(XX), and we have rationalized this result. The performance of the XDM dispersion model on the S22 database has been compared to a variety of alternative DFT methods that account for dispersion, and the XDM(BR) method compares very favorably. In future work, we will explore geometry optimizations of intermolecular complexes using XDM-derived dispersion forces.

**Acknowledgment.** The authors thank Takeshi Sato for a preprint of ref 33. This work is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Killam Trust of Dalhousie University. Computational facilities are provided by ACEnet, the Atlantic Computational Excellence network.

**Supporting Information Available:** Cartesian coordinates and reference binding energies for the training set of 65 intermolecular complexes (XYZ file format). This material is available free of charge via the Internet at http://pubs.acs.org/.

### References

(1) Kristyán, S.; Pulay, P. *Chem. Phys. Lett.* **1994**, *229*, 175.

(2) Johnson, E. R.; Mackie, I. D.; DiLabio, G. A. *J. Phys. Org. Chem.* **2009**, *22*, 1127.

(3) Andersson, Y.; Langreth, D. C.; Lundqvist, B. I. *Phys. Rev. Lett.* **1996**, *76*, 102.

(4) Dobson, J. F.; Dinte, B. P. *Phys. Rev. Lett.* **1996**, *76*, 1780.

(5) Kamiya, M.; Tsuneda, T.; Hirao, K. *J. Chem. Phys.* **2002**, *117*, 6010.

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

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

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

(9) Langreth, D. C.; Lundqvist, B. I.; Chakarova-Kack, S. D.; Cooper, V. R.; Dion, M.; Hyldgaard, P.; Kelkkanen, A.; Kleis, J.; Kong, L.; Li, S.; Moses, P. G.; Murray, E.; Puzder, A.; Rydberg, H.; Schroder, E.; Thonhauser, T. *J. Phys.: Condens. Matter* **2009**, *21*, 084203.

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

(11) Gerber, I. C.; Ángyán, J. G. *J. Chem. Phys.* **2007**, *126*, 044103.

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

(13) Goll, E.; Leininger, T.; Manby, F. R.; Mitrushchenkov, A.; Werner, H.-J.; Stoll, H. *Phys. Chem. Chem. Phys.* **2008**, *10*, 3353.

(14) Xu, X.; Goddard, W. A *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 2673.

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

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

(17) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 289.

(18) Sherrill, C. D.; Takatani, T.; Hohenstein, E. G. *J. Phys. Chem. A* **2009**, *113*, 10146.

(19) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. C* **2008**, *112*, 4061.

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

(21) Wu, X.; Vargas, M. C.; Nayak, S.; Lotrich, V.; Scoles, G. *J. Chem. Phys.* **2001**, *115*, 8748.

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

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

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

(25) Ortmann, F.; Bechstedt, F.; Schmidt, W. G. *Phys. Rev. B* **2006**, *73*, 205101.

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

(27) Chai, J.-D.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615.

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

(29) Lin, I.-C.; Coutinho-Neto, M. D.; Felsenheimer, C.; von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U. *Phys. Rev. B* **2007**, *75*, 205131.

(30) Sun, Y. Y.; Kim, Y.-H.; Lee, K.; Zhang, S. B. *J. Chem. Phys.* **2008**, *129*, 154102.

(31) DiLabio, G. A. *Chem. Phys. Lett.* **2008**, *455*, 348.

(32) Tkatchenko, A.; Scheffler, M. *Phys. Rev. Lett.* **2009**, *102*, 073005.

(33) Sato, T.; Nakai, H. *J. Chem. Phys.* **2009**, *131*, 224104.

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

(35) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2007**, *127*, 154108.

(36) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2005**, *123*, 154101.

(37) Ángyán, J. G. *J. Chem. Phys.* **2007**, *127*, 024108.

(38) Ayers, P. *J. Math. Chem.* **2009**, *46*, 86.

(39) Heβelmann, A. *J. Chem. Phys.* **2009**, *130*, 084104.

(40) Krishtal, A.; Vanommeslaeghe, K.; Olasz, A.; Veszpremi, T.; Van Alsenoy, C.; Geerlings, P. *J. Chem. Phys.* **2009**, *130*, 174101.

(41) Kong, J.; Gan, Z.; Proynov, E.; Freindorf, M.; Furlani, T. R. *Phys. Rev. A* **2009**, *79*, 042510.

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

(43) Johnson, E. R.; Becke, A. D. *J. Chem. Phys.* **2008**, *128*, 124105.

(44) Perdew, J. P. In *Electronic Structure of Solids '91*; Ziesche, P., Eschrig, H., Eds.; Akademie Verlag: Berlin, 1991.

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

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

(47) Becke, A. D. *J. Chem. Phys.* **1986**, *84*, 4524.

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

(49) Lacks, D. J.; Gordon, R. G. *Phys. Rev. A* **1993**, *47*, 4681.

(50) Kannemann, F. O.; Becke, A. D. *J. Chem. Theory Comput.* **2009**, *5*, 719.

(51) Perdew, J. P.; Yue, W. *Phys. Rev. B* **1986**, *33*, 8800.

(52) Perdew, J. P.; Yue, W. *Phys. Rev. B* **1989**, *40*, 3399.

(53) Murray, E. D.; Lee, K.; Langreth, D. C. *J. Chem. Theory Comput.* **2009**, *5*, 2754.

(54) Becke, A. D.; Roussel, M. R. *Phys. Rev. A* **1989**, *39*, 3761.

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

(56) Tang, K. T.; Toennies, J. P. *J. Chem. Phys.* **2003**, *118*, 4976.

(57) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 5656.

(58) Ruiz, E.; Salahub, D. R.; Vela, A. *J. Am. Chem. Soc.* **1995**, *117*, 1141.

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

(60) Chai, J.-D.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *128*, 084106.

(61) Becke, A. D.; Dickson, R. M. *J. Chem. Phys.* **1988**, *89*, 2993.

(62) Becke, A. D. *Int. J. Quantum Chem., Quantum Chem. Symp.* **1989**, *23*, 599.

(63) Becke, A. D.; Dickson, R. M. *J. Chem. Phys.* **1990**, *92*, 3610.

(64) Perdew, J. P.; Wang, Y. *Phys. Rev. B* **1992**, *45*, 13244.

(65) Handy, N. C.; Cohen, A. J. *Mol. Phys.* **2001**, *99*, 403.

(66) Becke, A. D. *J. Chem. Phys.* **2005**, *122*, 064101.

(67) Gulans, A.; Puska, M. J.; Nieminen, R. M. *Phys. Rev. B* **2009**, *79*, 201105.

(68) Klimes, J.; Bowler, D. R.; Michaelides, A. *J. Phys.: Condens. Matter* **2010**, *22*, 022201.

(69) Nilsson Lill, S. O. *J. Phys. Chem. A* **2009**, *113*, 10321.

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

(71) Schwabe, T.; Grimme, S. *Phys. Chem. Chem. Phys.* **2007**, *9*, 3397.

(72) Zhang, Y.; Yang, W. *Phys. Rev. Lett.* **1998**, *80*, 890.

(73) Mackie, I. D.; DiLabio, G. A. *J. Phys. Chem. A* **2008**, *112*, 10968.

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

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

(76) Wesolowski, T.; Parisel, O.; Ellinger, Y.; Weber, J. *J. Phys. Chem. A* **1997**, *101*, 7818.

(77) Zhang, Y.; Pan, W.; Yang, W. *J. Chem. Phys.* **1997**, *107*, 7921.

CT900699R

JCTC Journal of Chemical Theory and Computation

# A Modified Resonance-Theoretic Framework for Structure−Property Relationships in a Halochromic Oxonol Dye

Seth Olsen*

*Centre for Organic Photonics and Electronics, School of Mathematics and Physics, The University of Queensland, Brisbane QLD 4072 Australia*

Received January 1, 2010

**Abstract:** I demonstrate that a modification of the resonance color theory (in its form advocated by Brooker (*Rev. Mod. Phys.* **1942**, *14*, 275) and by Platt (*J. Chem. Phys.* **1956**, *25*, 80)) provides an accurate framework for rationalizing the *ab initio* excitation energies of the protonation states of the green fluorescent protein (GFP) chromophore (an asymmetric oxonol dye). I suggest that the original model space used in the resonance theory (specifically, a pair of Lewis structures) is formally inconsistent with a core aspect of the theory (specifically, a relationship between excitation energies and group-specific basicities (Brooker basicities) of the terminal rings). I argue that a more appropriate model space would consist of a complete active space ansatz based on group-localized orbitals. I then show that there is a solution to the state-averaged complete active space self consistent field (SA-CASSCF) problem with exactly this form. This family of SA-CASSCF solutions provides an objectively rigorous foundation for the resonance color theory. The solutions can be expressed in a localized set of active space orbitals, which display the same transferability pattern implied by the Brooker basicity scale. Using Platt's model Hamiltonian formulation of the resonance theory, I show that the accuracy of the set of excitation energies calculated with these solutions can be accurately reproduced using only two parameters per dye in the set. One of these parameters is the isoenergetic energy of the dye—the harmonic mean of the excitation energies of its symmetric parent dyes. The other parameter is a local basicity index (Brooker basicity), which is specific to each terminal ring and independent of the ring to which it is conjugated in a given dye. I proceed to show that the Brooker basicities, defined by differences between many-electron states, are also basicities in the usual (one-electron) sense and, finally, that Platt's construction of the color theory is an approximation to a *ab initio* effective Hamiltonian obtained by a minimum-norm block diagonalization procedure. What emerges is a powerful, simple, and *accurate* conceptual framework for thinking generally about color in monomethine dyes, and specifically about color tuning in the chromophore of green fluorescent proteins.

## Introduction

Understanding relationships between the optical properties of molecules and their chemical constitution is a long-standing goal of theoretical chemistry.[1] *Methine dyes*, such as those in Figure 1, have a prominent place in the history of these endeavors, and in chemistry itself. It is sometimes said that the modern organic chemical industry *began* with Perkin's discovery of the methine dye *Mauveine* in 1856, although the synthesis of the original *Cyanine* by Williams may predate this event.[2]

*Methine* is a limit of sp$^2$ carbon where the Lewis octet rule is satisfied by participation in multiple alternate struc-

* E-mail: s.olsen1@uq.edu.au. Website: http://www.uq.edu.au/∼uqsolse1.

10.1021/ct100001b © 2010 American Chemical Society
Published on Web 03/18/2010

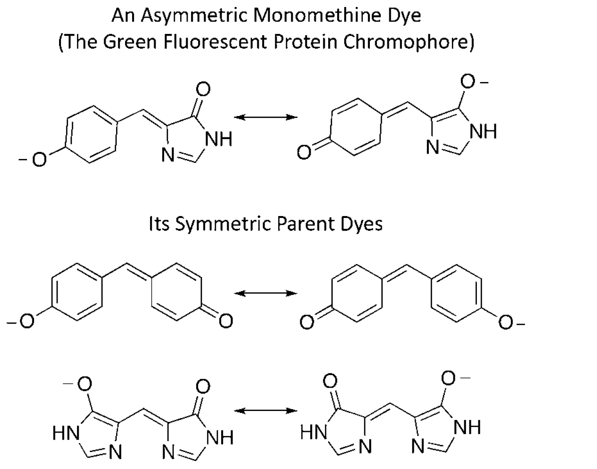

An Asymmetric Monomethine Dye
(The Green Fluorescent Protein Chromophore)
Its Symmetric Parent Dyes
O
-O
NH
N
O-
Ο
NH
N
HN

***Figure 1.*** The classical resonance theory of methine dye color relates the color of an asymmetric methine dye (such as the Green Fluorescent Protein chromophore, top) to the colors of its symmetric parents (bottom), and to the energy difference between the resonance forms for the asymmetric dye. The latter is dominated by the difference in electronic basicities of the terminal heterocycles, which bear the charge in different structures.

tures, which differ by bond alternation and formal charge relocation. An archetypical *methine dye* is composed of two groups, usually heterocyclic rings, each of which can access two redox states, separated by a bridge containing one or more methine units. A *monomethine dye* has a single methine unit in the bridge, while a *polymethine dye* has more than one.

When the terminal groups are different, one of the resonating structures will dominate over the other. Within this context, the resonant methinic structure is a particular limit in a continuous set, the extreme elements of which are ionic polyenic structures with definite and opposing bond alternation. Though the dyes in Figure 1 are ions, the methine electronic structure is also relevant to neutral donor−acceptor chromophores. Within these systems, which resonate between neutral and zwitterionic closed shell states (as opposed to diametrically opposed charge structures as in Figure 1), it has been shown that the optical properties of the chromophores are strongly dependent on their proximity to the methine limit.[3] For this reason, the methine electronic structure is a subject of continuuing interest in the study of organic photonic and electronic materials.[4–7]

The molecule that I use an example at the top of Figure 1 is the chromophore of the green fluorescent protein, a system which has lept to prominence[8] for its utility in biooptical technologies.[9] The chromophore is a *p*-hydroxybenzylidene-imidazolinone motif−an oxonol dye system. The chromophores of almost all fluorescent proteins are derivatives of this dye structure.[10] Many of the technologies for which fluorescent proteins are used depend in some way on the color of the protein,[9] which in turn depends on the color of the chromophore. The color of proteins with derived chromophores cluster together according to the nature of the derivation.[10] Spectral variation within each group derives from variations in the interaction of the chromophore with its protein.[11] Each fluorescent protein chromophore is synthesized within a particular protein and bound inside that protein for the duration of its functional existence.[10] Fluorescent proteins are therefore very interesting natural laboratories for studying structure−property and environment−property relationships in methine dyes.

The issue of the color of fluorescent protein chromophores is a good example of how the general problem of color and constitution in methine dyes has maintained its relevance, despite its already long history.[12]

## The Resonance Color Theory and the Brooker Basicity Scale

The problem of color and constitution in methine dyes received considerable attention during the early part of the twentieth century.[1,13–19] Considerable motivation for development derived from the herculean experimental program of LGS Brooker and his group at Kodak, where a great many methine dyes were synthesized and studied (methine dyes are useful photographic sensitizers).[18] Methine dyes had been known for almost a century when Brooker's work emerged, so his group was far from the first to study these systems. However, the scale and scope of the contribution have raised Brooker's name to prominence. For a thorough review of early work, I recommend a crucial (and very well-referenced) paper by Platt.[14] I also recommend the books by Griffiths[2] and by Fabian and Hartmann,[20] as well as an article by Berneth in *Ullmann's Encyclopedia of Industrial Chemistry*.[21] An early work of particular interest, which Platt's work builds upon, is Kuhn's description of the methine limit as a free-electron gas.[13] Few theories can rival it for simple, effective insight−save possibly Platt's perimeter model of cata-condensed hydrocarbons, published in the same year.[22]

The spectroscopy of methine dyes up to and including Brooker's studies had established several empirical rules,[14] one of which will be our particular focus. The rule in question relates the color of an asymmetric methine dye (such as the GFP chromophore in Figure 1) to the color of the symmetric "parent" dyes produced by conjugating each terminus with a copy of itself (for the GFP chromophore, such dyes are shown at the bottom of Figure 1). This rule−*The Deviation Rule*[18]−is summarized as follows:

*The absorbance wavelength of an asymmetric dye is no redder than the mean wavelength of its symmetric parents and deviates from this by a blue shift which increases as the difference in basicity of the terminal groups.*

It is clear from Brooker's language[17,18] that, when he used the word "basicity", he was using it in the sense that it is normally meant−an energy associated with a *one-electron* process. This interpretation is implicit in theoretical works that followed Brooker's papers−specifically Herzfeld and Sklar's tight-binding Hamiltonian treatment[23] and Kuhn's free-electron model,[13] both of which describe the chemical asymmetry in terms of a one-electron potential.

Brooker measured the absorption of many dyes and catalogued the deviations ("Brooker deviations") of these from the mean of each dye's symmetric parents' parents.[18] He used this data for formulation of an affine basicity scale for the terminal groups in the dyes ("the Brooker basicity scale").[18] The Brooker basicity scale is correlated with other basicity scales such as the Hammet $\sigma_R$ scale.[2] Platt's contribution was to verify that the deviations in Brooker's data set could indeed, to within the experimental error, be expressed by two numerical parameters: one specifying the mean parental wavelength and the other a basicity index specific to each terminal group, *and independent of the group to which it was conjugated a given dye.*[14]

In rationalizing his results, Brooker invoked a theory of the color of dyes, which was framed in a heuristic model space of resonating Lewis structures.[17,19] Within this theory, the optical excitation of the dye emerges from the effective coupling between two isoenergetic *extreme* structures, which are analogues of those in Figure 1.[16] Overlap arguments suggest that the direct coupling will be too small to explain the commonly measured wavelengths, so the theory hypothesizes *intermediate* structures, which place the formal charge on the bridge.[16,19] The latter postulate is particularly important in long polymethine chains, where the exponential decay will decimate the bare coupling.[24] The intermediate structures are higher in energy than the extreme structures. They play an indirect role in the optical excitation, for which purpose they can be expressed as an effective potential.[16,24] This potential creates a gap between the extreme structures. The excitation should be optically intense by very simple dipole length arguments.[25,26]

There is only one bridge, and since the electronic structure on the bridge differs only by bond alternation in the extreme structures, the residual splitting between the extreme structures will be dominated by the difference in basicity of the terminal groups.[18] This residual splitting is present in the absence of interaction and leads to a blue shift of the dye relative to its symmetric parents.[14,16] This implies a two-state picture where the interaction matrix element is equal to the harmonic mean (in energy units, the arithmetic mean in wavelength) of the excitation energies of the symmetric parent dyes.[14]

Although the resonance theory in its classical form generated very effective heuristic explanations, attempts to translate it into quantitative models were problematic.[14,16] It appears the resonance color theory faded into obscurity, a casualty of the early competition between molecular orbital (MO) theories of electronic structure and the valence-bond (VB) theories that were descendents of the early resonance theories.[27–29] MO representations offer efficient techniques for storing and manipulating many-body states in the Born−Oppenheimer (clamped classical nuclei) electronic structure problem, because the one-electron density operator matrix elements span a Lie algebra.[30,31] Valence bond theories have other strengths, particularly for constructing diabatic states whose character is maintained over an open neighborhood of nuclear geometries, and which can more easily accommodate nuclear motion and bond rearrangements.[32,33] Their representation is less economical, so a computational threshold had to be crossed before they were incorporated into regular computational use.[34,35] It is now well-known that MO and VB are only different ways of generating bases for the quantum mechanical state space.[36,37] States are vectors in quantum mechanics, so the representation does not matter if both representations can be spanned within the same *complete* space.[38] Observables in both representations can be represented in the same algebra.[39–41]

## Platt's Construction

Platt constructed a quantitative empirical framework for the resonance color theory.[14] Platt's construction is a recipe for generating 2 × 2 model Hamiltonians for a set of dyes built from a common set of terminal groups. The construction describes a given dye with two parameters, to be extracted from empirical data. The first of these is the "isoenergetic excitation energy", which is the harmonic mean of the excitation energies of the parent symmetric dyes (proportional to the mean wavelength).

$$\frac{1}{E_I(\mathrm{A,B})} = \frac{1}{2}\left(\frac{1}{E_I(\mathrm{A,A})} + \frac{1}{E_I(\mathrm{B,B})}\right) \tag{1}$$

Here, $E_I$(A,B) is the isoenergetic excitation of the dye (A,B) generated with terminal groups A and B. The isoenergetic excitation of a symmetric dye (such as (A,A) and (B,B)) is equal to its excitation energy. The second quantity in the Platt construction is the Brooker basicity difference $b$(A,B).

$$b(A,B) = \sqrt{(\Delta E(A,B))^2 - (E_I(A,B))^2} \tag{2}$$

where $\Delta E$(A,B) gives the excitation energy of the dye (A,B) generated with terminal groups A and B. Platt's construction yields a traceless model Hamiltonian $H^{\mathrm{P}}$(A,B).

$$H^{\mathrm{P}}(A,B) = \frac{1}{2}\begin{pmatrix} b(A,B) & E_I(A,B) \\ E_I(A,B) & -b(A,B) \end{pmatrix} \tag{3}$$

The definition of $b$(A,B) ensures that the splitting between the eigenvalues of $H^{\mathrm{P}}$(A,B) is equal to the dye excitation energy $\Delta E$(A,B).

Platt's primary contribution was to show that the quantity $b$(A,B) could be expressed as a difference between basicities that were *constant for each terminal group in the set*. He showed that Brooker's data could, within the experimental error, be described by a $b$(A,B) formula with the simple parametric form (eq 4).

$$b(\mathrm{A,B}) = b_{\mathrm{A}} - b_{\mathrm{B}} \tag{4}$$

where $b_{\mathrm{A}}$ and $b_{\mathrm{B}}$ are *constants* characteristic of groups A and B *and are independent of the conjugate groups with which they paired in any given dye*. Platt showed this by demonstrating that the $b$(A,B)'s extracted from Brooker's data obey the following "consistency rules".

$$b(\mathrm{A,C}) - b(\mathrm{B,C}) = b(\mathrm{A,D}) - b(\mathrm{B,D}) \tag{5}$$

$$b(\mathrm{A,C}) = b(\mathrm{A,B}) + b(\mathrm{B,C}) \tag{6}$$

The consistency rules above are not actually independent, since the second can be derived from the first, provided that the $b$(A,B) actually can be written as a difference (so that $b$(A,B) = $-b$(B,A)).

Platt went on to demonstrate that a large data set published by Brooker could be compactly summarized and reproduced by his construction, using a suitable set of $b_{\mathrm{A}}$ parameters.[14]

In what follows, I will demonstrate that the excitation energies obtained for an example set of dyes within an *ab initio* representation that mimics the structure of the model space in color theory *can also be expressed this way*. That is, I will extract a set of $b_A$'s from a set of calculated excitations, and show that the excitation reconstructed with these parameters does not meaningfully deviate from the input set. I will then demonstrate that the $b_A$'s actually do measure a one-electron energy difference and that the Platt construction can be considered as a synthetic approximation to a quasi-diabatic *ab initio* effective Hamiltonian.

## Revising the Model Space in the Resonance Color Theory

I want to point out that the information content implied by the Lewis structural representation of resonance theory, outlined in Figure 1, *is not formally consistent* with the definition of the Brooker basicity scale. There are different ways to highlight the problem. The problem is that a "basicity" usually means an energy associated with a *one-electron* process, while the Lewis structures in Figure 1 are rich in *pair* information (i.e., the bonding). Operators on a one-electron Hilbert space will generally not commute with pair operators defined on the tensor product of the space with itself. Although the set of one-electron operators spans a Lie algebra closed under commutation, the set composed of one- and two-electron operators generally does not.[30] This means that not only do the one- and two-electron operators not commute but expressing the commutator requires expanding the set of operators.[42] *These statements imply that uncertainty relations prevent the precise, simultaneous specification of the basicity of a group and the bonding within the group*. A second argument simply notes that the underlying one-electron basis implied in the Lewis structure representation is one of the (perhaps orthogonalized) atomic orbitals. If the basicities of the different states in the group are not equal, then there will be *multiple detachment/attachment states, and a distribution of possible basicity values, associated with the ring*. This will be true even for a single dye molecule in a given Lewis structure. These problems are already apparent in discussions put forward by Brooker, who, *as expected*, had to invoke additional resonance structures beyond the canonical pair (e.g., Figure 1) for the dyes that he studied in order to rationalize his results.[18]

I propose that a more appropriate model space for the formulation of the resonance color theory could be obtained from a "methine-adapted" complete active space valence bond (CASVB) ansatz[38] built from group-localized orbitals, such as that outlined in Figure 2. This ansatz is consistent with the precise definition of group basicities, if the basicities are defined so that the group-localized orbitals are the relevant attachment/detachment states. Furthermore, the ansatz in Figure 2 still contains enough *pairing* degrees of freedom to index states in the model space spanned by the canonical resonating structures (Figure 1). This means that the information content of the theory can be preserved.

In molecular orbital theories, monomethine dyes are related to *odd alternate* systems.[15,43] This means that, when the $\pi$ molecular orbitals are paired according to their bonding or antibonding character, there will be a nonbonding orbital (NBO) left over. In diarylmethine dyes (for example Michler's hydrol blue), the NBO is doubly occupied. In monomethine cyanines (for example Williams' cyanine), this orbital is empty. In MO theories, the optical excitation of the former class is a HOMO−LUMO excitation from the occupied NBO to the lowest antibonding orbital. In the latter class, the excitation is a HOMO excitation from the highest bonding orbital into the NBO. The symmetry of odd alternant hydrocarbons in simple MO theories is such that these excitations would be the same for the anion and the cation formed from the same molecular frame. This implies that the four-electron/three-orbital CASVB representation outlined in Figure 2 unifies the state spaces of the resonance color theory and odd alternate MO theories for diarylmethines, and the corresponding two-electron/three-orbital CASVB space does the same for the monomethine cyanines. The GFP chromophore system is not strictly alternate, because it contains rings with an odd number of sites. Even so, it is *isoelectronic* with an odd-alternate hydrocarbon, and its orbitals can be identified with such a system. *Therefore, a model space over three frontier orbitals is also indicated by simple MO theories.*

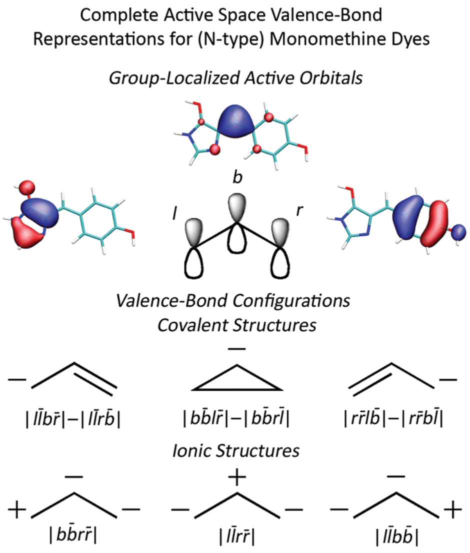

Complete Active Space Valence-Bond
Representations for (N-type) Monomethine Dyes
Group-Localized Active Orbitals
b
l
r
Valence-Bond Configurations
Covalent Structures
$|l\bar{l}b\bar{r}|-|l\bar{l}r\bar{b}|$
$|b\bar{b}l\bar{r}|-|b\bar{b}r\bar{l}|$
$|r\bar{r}l\bar{b}|-|r\bar{r}b\bar{l}|$
Ionic Structures
$|b\bar{b}r\bar{r}|$
$|l\bar{l}r\bar{r}|$
$|l\bar{l}b\bar{b}|$

***Figure 2.*** A model space for the resonance theory that is consistent with the precise basicity value for the rings would be a "methine-adapted" complete active space valence bond (CASVB) representation with one orbital state for each of the terminal rings (plus one for the bridge). The Lewis structure-based model space is not consistent with a precisely defined basicity, because there is too much information needed to specify the rings' internal structure. As a result, given a dye in a precisely defined Lewis structure, there would be a spread of possible basicities.

## A Self-Consistent Representation of the Resonance Color Theory

There is a methine-adapted solution to the two-state-averaged[44,45] complete active space self-consistent field[46,47] (SA2-CASSCF) problem with the form of Figure 2 for a

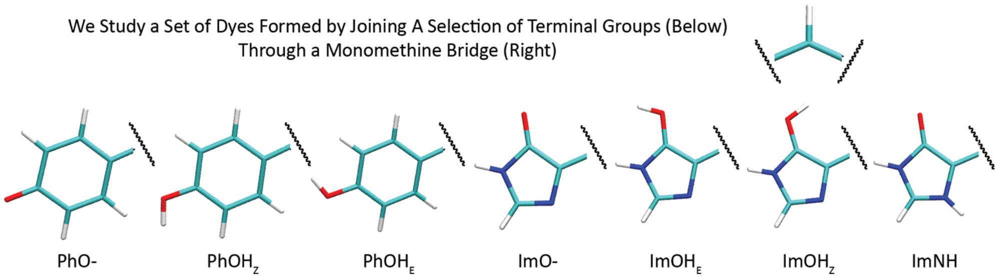

We Study a Set of Dyes Formed by Joining A Selection of Terminal Groups (Below)
Through a Monomethine Bridge (Right)
PhO-
PhOH_Z
PhOH_E
ImO-
ImOH_E
ImOH_Z
ImNH

***Figure 3.*** The data set used here consists of excitation energy calculations on a set of monomethine dyes generated by conjugating phenoxy and imidazolinoxy groups in different titration states (shown here). The generated dye set includes several conceivable protonation states of the green fluoroescent protein chromophore, as well as several *bis*-phenoxy and *bis*-imidazolinoxy dyes.

large range of monomethine dyes. This includes the systems I use here as examples but appears to be much more general.[48] I conjecture that the existence and stability of such a solution can be safely used as an *operational definition* of the term "monomethine dye". In a complete active space representation, the dyes I examine here have an active space with four electrons in three orbitals (SA2-CAS(4,3)). The solutions obtainable for monomethine cyanines have two electrons in three orbitals (SA2-CAS(2,3)).[49] In either case, the orbitals, after localization of the active orbitals with the Foster−Boys technique,[50] have the group-localized structure shown in Figure 2. The Boys localization procedure is unitary, so that the CASSCF state is invariant as long as the transformation respects the boundaries between the occupied, active and virtual spaces.[45,51]

I argue that the methine-adapted SA-CASSCF solutions described above are *self-consistent representations* of the resonance color theory. This is important, because it highlights a strategy for making quantitative and objective predictions where the old resonance theory yielded only heuristic ones. It also means that the resonance color theory rests on stronger foundations than had previously been apparent.[2] The domain of application of the theory can be objectively assessed as the domain of applicability of the corresponding self-consistent field solutions.

## A Set of Example Dyes

In what follows, I will support my assertion of correspondence between the resonance color theory. To do this, I will first describe a data set of excitation energies obtained by multistate[52] multireference second-order perturbation theory[53,54] (MS-MRPT2) on the methine-adapted SA-CASSCF reference space. Then, I will apply Platt's synthetic Hamiltonian construction[14] to this data set. Finally, I will show that *the accuracy of the excitation energies recalculated using Platt's construction is within the expected accuracy of the original calculations themselves*. I will use a data set composed of 28 independent dye structures, all of which are monomethine pairings of different protonation states of a phenoxy and imidazolinoxy group (e.g., Figure 3). This set includes several protonation states of the GFP chromophore motif (which have been previously studied by quantum chemistry[55−83]).

In general, methine dyes can have multiple *cis*−*trans* isomeric states. In the context of the example set here, there is no meaningful distinction between *cis* and *trans* isomeric states of the phenoxy−methine bond (due to symmetry about its axis). Such a distinction is meaningful only for the imidazolinoxy−methine bond, for which I examine only the *cis* forms here (the imine nitrogen at position 2 is *Z* with respect to the conjugate terminal group). Each dye was relaxed in its *cis* conformation by performing an MP2[84] optimization with a cc-pvdz basis set.[85] For a few dyes in the set, the ground state minimum is a different *cis*−*trans* isomer and is not contained in the set. As my goal is to investigate the resonance color theory as a theory of the *electrons*, restricting the set of structures in this way makes sense. This procedure generates structures that are minima with respect to bond alternation coordinates. This is important, because one would expect that the model states of the resonance theory (e.g., Figure 1) are coupled through bridge bond-stretching vibrations.

For each dye in its relaxed *cis* geometry, I obtained the "methine adapted" SA2-CAS(4,3)/cc-pvdz solution space using unrestricted Hartree−Fock charge-density natural orbitals[86,87] for the oxidized doublet radical as an initial guess for the self-consistent field optimization. I then calculated the excitation energy of the dye by applying a multistate multireference perturbation theory (MS-MRPT2) correction to the SA2-CAS(4,3) reference space. The MS-MRPT2 correction is formally size-extensive.[88] This is consistent with the interpretation of the underlying CASSCF as a form of maximum entropy inference.[31,89] Only the highest-lying 32 orbitals were correlated, though, and some extensivity error might arise from this.[90] The perturbation theory calculations on the methine adapted SA2-CAS(4,3) solutions converged quickly and easily without the use of level shifts. The above procedure yielded 28 excitation energies for symmetric and asymmetric dyes generated by the groups in Figure 3. Additional details, as necessary to reproduce the wave function (i.e., state-averaged natural orbitals and occupation numbers, and MS-MRPT2 mixing matricies) are available in the Supporting Information. The most concrete result of this paper is that the resulting set of excitation energies can be accurately represented by Platt's construction,[14] where each group is assigned its own basicity index independent of the terminal group to which it is conjugated in a given dye.

I used the Molpro software package for all electronic structure computations.[91]

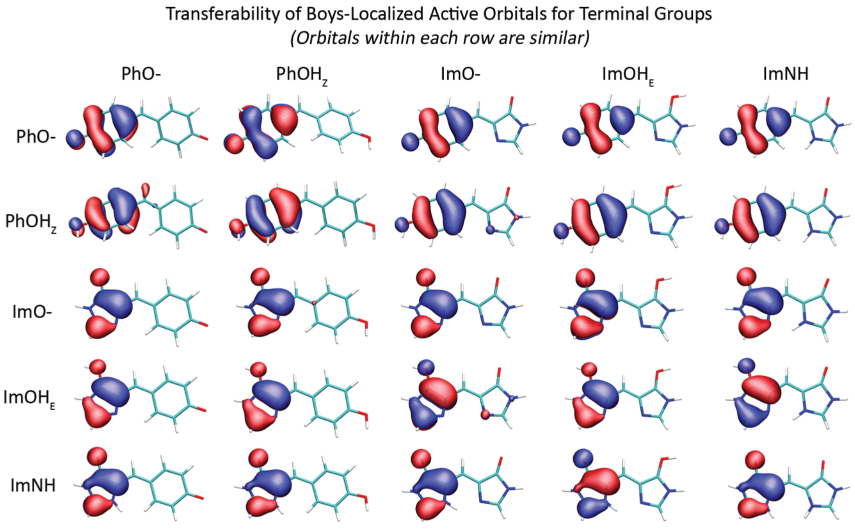

Transferability of Boys-Localized Active Orbitals for Terminal Groups
(Orbitals within each row are similar)
PhO-
PhOH$_Z$
ImO-
ImOH$_E$
ImNH
PhO-
PhOH$_Z$
ImO-
ImOH$_E$
ImNH

***Figure 4.*** The Boys-localized active space orbitals obtained from the methine-adapted SA2-CAS(4,3) solution for the example dyes are transferrable in the manner suggested by the Brooker basicity scale. The shape of the orbital for each group is independent of the group to which it is paired in a given dye structure (i.e., orbitals within each row are similar). The Brooker scale associates a basicity (one-electron energy) to each group, independent of its context. Orbital isosurfaces are set at ±0.04. Overall, signs do not matter, by the SA-CASSCF convergence criteria.

My goal is to show that the methine-adapted solution to the SA-CASSCF problem *contains* the resonance color theory. We are now at a point where I can present the first pieces of suggestive evidence in defense of this notion. These are highlighted in Figures 4 and 5. Figure 4 shows that the Boys localized active space orbitals associated to each of the rings in the set of dyes maintain their shape when the conjugate dye is varied. This is exactly the pattern expected in order for a transferrable basicity scale—such as the Brooker scale—to apply.[18] Figure 5 shows that the excitation energies obtained for the asymmetric dyes are consistent with Brooker's deviation rule, in that none of the asymmetric dyes shown is redder than the mean wavelength of its parents.[92]

I want to highlight a particular structure to the excitation energies generated using the SA2-CAS(4,3) ansatz, not necessarily to argue for its absolute accuracy as an approximation to the exact Born−Oppenheimer electronic structure. It is convienent, however, that the excitations *do* compare well against observed excitations for chemically similar systems. In particular, the excitation energy of the anionic phenolate−imidazolinolate (PhO−, ImO−) dye is in reasonable range of the measured excitation energies of green fluorescent protein (GFP) chromophore models in their anionic state,[93−97] as well as the B band of GFPs,[98−100] and is broadly consistent with computational models using larger active spaces.[60,63,72,75,82] Moreover, the excitation energy of the bis-phenolate (PhO−, PhO−) dye is quite close to the lowest excitation of benzaurin and phenolpthaleins in alkaline solution, and the corresponding diprotonated cations (PhOH/PhOH) are close to the excitation of the benzaurin cation.[2] It does seem, interestingly, that there may be a systematic overprediction of the excitation energies of the neutral oxonol dyes in the set (e.g., (PhOH, ImO−), (PhO−, ImOH), etc.). The calculated excitations of the (PhOH, ImO−) and (PhO−, ImOH) dyes are both bluer than GFP chromophore models at neutral pH in several solvents,[96] models in an ion ring,[55] the A band of GFPs,[98−100] and results from larger active spaces.[55,83] Similarly, the excitation of the (PhOH, PhO−) dye is bluer than the absorbance of benzaurin at neutral pH.[2] It may be that the estimates provided by the methine-adapted solution spaces overestimate the blue shift near the polyenic limits of the resonance scale. This question is not relevant to my purpose, which is to show that the excitations *within this model* follow a specific simple pattern, *and this pattern is predicted by the resonance color theory*. The difficulty of producing quantitatively accurate absolute excitation energies for dyes contained in the set has been highlighted in two recent benchmarking studies using similar techniques.[60,61]

It is worth noting that the dimethyl derivative of the (PhO−, ImO−) dye (HBDI), is autoionizing in its first excited state.[60] This is interesting, because one might expect this to artificially depress the excitation energy, and violate Brooker's deviation rule. Apparently, this is not happening (Figure 5). Possible explanations may be that (*a*) compensating artifacts occur in the parent dyes, so that the deviation rule is preserved, (*b*) the dye is nonresonant but appears so due to the artifact, (*c*) the calculation is (somehow[101]) managing to pick out the appropriate valence state from the embedding continuum, or (*d*) substituent effects induced by removing the methyl groups raise the ionization threshold above the first excited state.

## Extraction and Validation of Basicity Indices for Terminal Groups

I extracted single basicity indices for each terminus in the following steps. First, I collected all basicity differences *b*(A,C) − *b*(B,C) using Platt's construction,[14] where A, B, and C ranged over the set of termini (Figure 3). I grouped these according to B and performed a linear regression fit within each of these groups. This step yielded the data in Figure 6. The lines for each B are parallel to a very good approximation. Therefore, the scales given by different B groups can be expressed relative to a common origin by

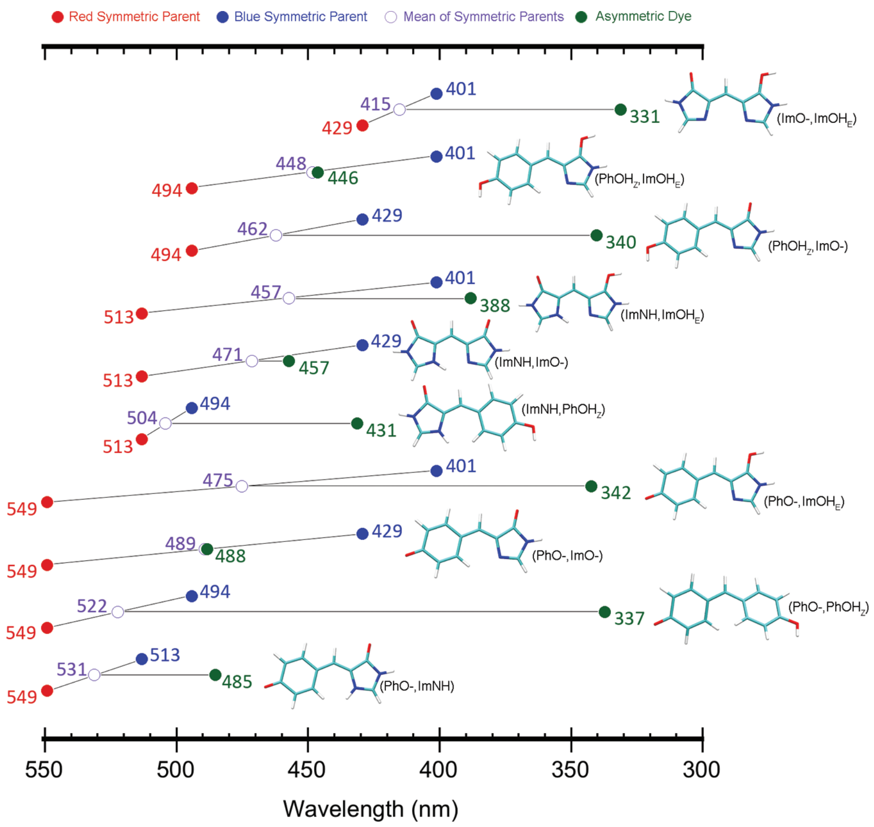


***Figure 5.*** Excitation wavelengths calculated using the methine-adapted SA2-CAS(4,3) solutions are consistent with the Brooker deviation rule. Calculated excitation energies for a selection of dyes in the data set are shown. Excitations were calculated using MS-MRPT2 theory corrections to the methine-adapted SA2-CAS(4,3)/cc-pvdz reference model spaces. Excitations of asymmetric dyes are green dots and type, excitations of symmetric parents are red and blue dots and type, and mean wavelengths of the symmetric parents are purple circles and type. To the right of each row is a picture of the corresponding asymmetric dye, and a label. All of the calculated excitations are consistent with Brooker's deviation rule, because the excitation wavelengths of the asymmetric dyes are not redder than the mean wavelengths of their symmetric parents. Brooker deviations are greatest when the titration states of the aryloxy sites are different, indicating strong detuning from resonance. When they are the same, as in the anion (PhO−, ImO−) and dihydroxy cation (PhOH$_Z$, ImOH$_Z$), the deviation is small, indicating resonance.

shifting each group according to the *y* intercept of the respective regression line. This was also supported by the application of Friedman's test[102] to the data, which indicated a high degree of agreement between scales corresponding to different choices of B.

After shifting, the basicities corresponding to each A are clustered together. Figure 7 shows a histogram of the shifted basicities colored according to the value of A.

I extracted a single basicity for each terminal group A by taking the median of the distribution of shifted basicities for each A. The median basicities and associated median absolute deviations are listed in Table 1.

If the resonance theory is a reasonable model for the excitations in the data set, then reconstructing the excitation energies for each of the asymmetric dyes in the set using Platt's construction[14] should not significantly degrade the accuracy of the set. I tested this by back-calculating the excitation energies for each asymmetric dye in the set using the median shifted basicities and the isoenergetic excitations of the parent symmetric dyes. The set of reconstructed excitation energies fits the original set very well, as I show in Figure 8, by a linear fit between the two sets. The residuals of the fit (vertical distances to the regression line) were under 1000 cm$^{-1}$ (0.123 eV) for all of the dyes in the set. *This is as good a level of accuracy as would be expected a priori from quantum chemical estimates anyway,*[60,75,103–109] *so the approximation of using group-specific basicities does not meaninfully degrade the accuracy of the set.*[110]

## Brooker Basicities are Correlated with One-Electron Basicities

An interesting feature of the resonance color theory is that the "Brooker basicity" is defined by a difference between observables on a *many-electron* Hilbert space. Yet, it is interpreted as a real chemical "basicity", which usually implies a *one-electron* energy. Even more interesting is that this relationship is experimentally verifiable.[2] I will now show that this relationship also emerges in the *ab initio* SA2-CAS(4,3) representation, by showing that the Brooker (many-electron) basicity scale extracted from the calculated excitations is correlated with a one-electron basicity scale extracted from the set of Boys-localized active space orbital energies (diagonal state-averaged Fock matrix elements).

I extracted group-specific one-electron basicities for the terminal groups in much the same way as for the Brooker basicities. Specifically, I collected differences $b'(A,C) - b'(B,C)$ where $b'(A,C)$ is the difference between the orbital

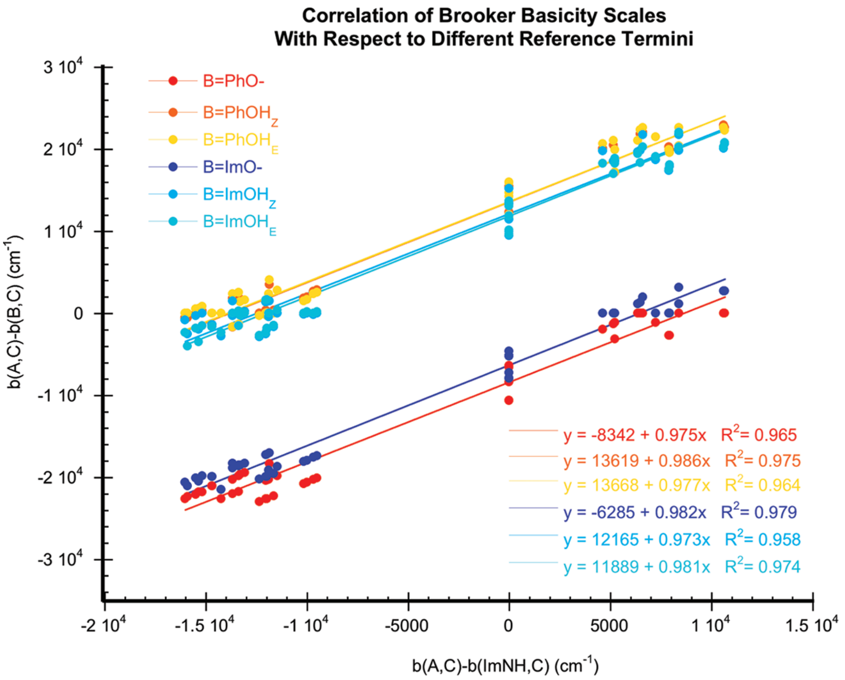


***Figure 6.*** The Brooker basicities assigned to the different termini are consistent, if any single termini is chosen as a reference for the scale. The basicities with respect to the ImNH terminus are used as the abcissa, and basicities with respect to other termini are plotted against these. Relative Brooker basicity differences $b(\text{A,C}) - b(\text{B,C})$ are plotted for all choices of groups B and C (see Figure 3). Groups corresponding to different B's were fit to a linear regression model. Coefficients of determination ($R^2$) are shown, indicating that each group B $\neq$ ImNH is well correlated with B = ImNH. The lines fit to different B's are parallel with a slope close to unity, indicating that they are also well correlated with each other. Groups corresponding to different choices of B can be brought into the same scale by shifting each according to the *y* intercept of its regression line.

energies (diagonal state-averaged Fock matrix elements) of Boys-localized active orbitals corresponding to the groups A and C in the dye (A,C). I then followed the same procedure that I used to extract the Brooker basicities $b_A$ from the differences $b(\text{A,C}) - b(\text{B,C})$ above. The data behaved in a very similar fashion to the Brooker deviation data at each step (Figures S1.1 and S2.2 in the Supporting Information are analogs to Figures 6 and 7 for the Brooker data). After extracting the group-specific one-electron basicities, I performed a linear regression fit between the Brooker and Boys orbital basicities to determine their correlation. I show the results of this comparison in Figure 9. The Brooker and Boys basicities are strongly correlated. This demonstrates that the *ab initio* Brooker basicity does measure a group "basicity" in the usual (one-electron) sense. *This is a core assertion of the resonance color theory.*[14,18]

The relationship between the many-electron (Brooker) and one-electron basicity scales should reflect the importance of electron correlation in the electronic structures of the terminal groups. Specifically, the distance from the (*one-* vs *many-electron*) regression line should reflect the relative importance of electronic correlations to the basicities. With this in mind, Figure 9 suggests that electron correlations are most important for the oxygen-protonated termini $\text{PhOH}_{E/Z}$ and $\text{ImOH}_{E/Z}$. The orderings of basicities of the $\text{PhOH}_{E/Z}$ and $\text{ImOH}_{E/Z}$ termini are apparently reversed between the one-electron and many-electron scales. The difference to the regression line is still small compared to the total range sampled, so it is probably reasonable to say that the basicities of the $\text{PhOH}_{E/Z}$ and $\text{ImOH}_{E/Z}$ termini are not operationally distinguishable. This viewpoint was supported by a statistical variance analysis of the entire distribution of shifted basicities for all terminal groups in the set, wherein the distinguishability of these two specific groups depended sensitively on the parameters used in the test. These tests also indicated that the distributions corresponding to $\text{OH}_E$ vs $\text{OH}_Z$ conformations were indistinguishable, so that the basicity does not depend on the oxygen lone pair to which the proton is bound.

## Platt's Construction Approximates an Ab Initio Effective Hamiltonian

Platt's construction[14] is a recipe for *synthesizing* a 2 × 2 spectroscopic Hamiltonian from a collection of given absorption wavelengths. Though Platt's language[14] suggests he had the Lewis structural representation in mind, he did not actually write down any constraints on the form of the representation to which his synthetic Hamiltonian matrix corresponds. The representation is defined only through the matrix elements of the Hamiltonian he constructed. It is safe to conclude that *any* representation that obeys the right relationships between the absorbance wavelengths of a collection of asymmetric dyes and their symmetric parents, and for which a consistent set of basicities can be defined, is a candidate. One could, if one wished, consider an ensemble of representations consistent with the constraints (perhaps supplemented by other physically motivated constraints) and consider the state as a random variable.[111] In the context of the modified resonance-theoretic model space that I have proposed, it seems reasonable to insist that the constraints include a map between the energies defined in

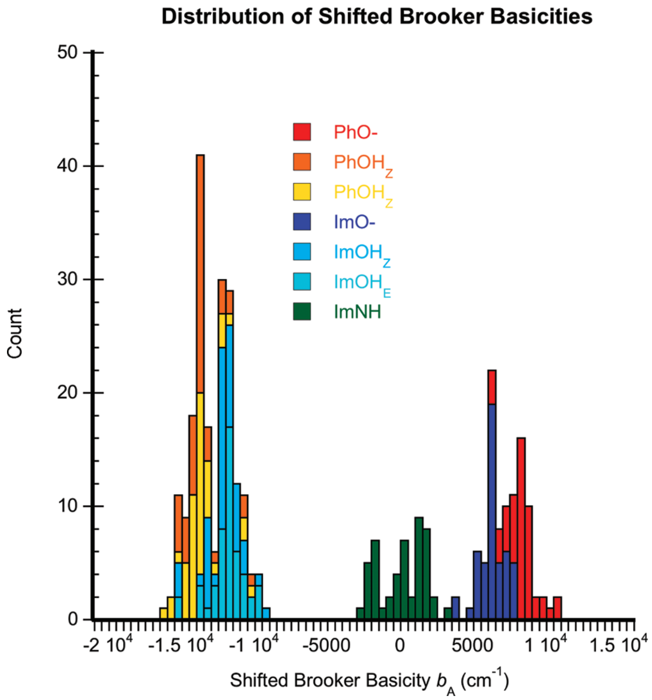


***Figure 7.*** This figure shows the distribution of group-specific Brooker basicities generated when the relative basicity differences $b(A,C) - b(B,C)$ are shifted onto the same line (see Figure 6). This yields tightly clustered distributions for each group A, containing 25 data points each. The distribution for each A is sharply peaked and, for most cases, is concentrated within 5000 cm$^{-1}$ of the peak. This suggests that the excitation data may be summarized without too much loss of accuracy using Platt's model Hamiltonian construction with one median Brooker basicity parameter assigned to each group.

***Table 1.*** Median Brooker Basicities (cm$^{-1}$) for Different Terminal Groups in the Data Set and the Median Absolute Deviations of the Distributions for Each Terminal Group

| group | median $b_A$ (cm$^{-1}$) | median absolute deviation (cm$^{-1}$) |
|---|---|---|
| PhO− | 8342 | 654 |
| PhOH$_Z$ | −13658 | 523 |
| PhOH$_E$ | −13668 | 576 |
| ImO− | 6288 | 614 |
| ImOH$_Z$ | −12165 | 664 |
| ImOH$_E$ | −11889 | 566 |
| ImNH | 0 | 1272 |

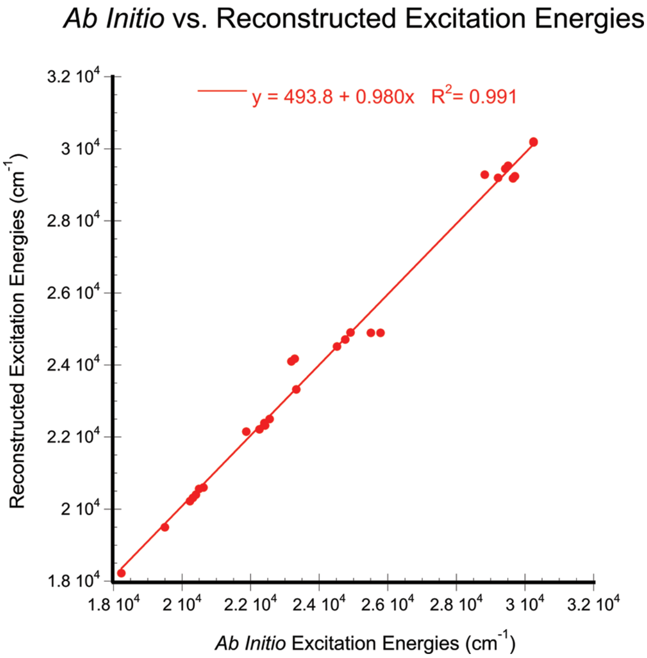


***Figure 8.*** Reconstruction of the excitation energies in the data set using median Brooker basicity values (rather than the dye-specific values) does not meaningfully degrade the set. The calculated and reconstructed excitations (cm$^{-1}$) for dyes in the set are plotted against one another and fit to a linear regression model with parameters displayed. The coefficient of determination ($R^2$) is close to unity, showing that the data are strongly correlated. The slope of the regression line is also very close to unity, as expected. All of the residuals (vertical distances from the regression line) are less than 1000 cm$^{-1}$. This is much smaller than the range spanned by the set (~10 000 cm$^{-1}$) and comparable to the best accuracy expected for correlated quantum chemical excitation energy estimates. Therefore, the use of group-specific Brooker basicities summarizes the data set compactly and *with no significant loss of accuracy*.

the one- and many-electron spaces, so that the Brooker basicities measure basicities calculated with the one-electron set.

I have shown above that the family of methine-adapted SA2-CAS(4,3) solutions for the example dye set obeys these requirements, to within the expected accuracy of the computations. We can use this to probe the relationship between the Platt Hamiltonian and other 2 × 2 Hamiltonians that can be extracted from the quantum chemical model space. For example, I could extract the angle between the Platt Hamiltonian and the Hamiltonian defined in the eigen representation of the SA-CASSCF solution (or its image under rotation by the MS-MRPT2 mixing matrix). Another interesting candidate for comparison would be the 2 × 2 Hamiltonian obtained by a *minimum-norm block diagonalization transformation*.[112] This is the transformation that does as little as possible other than block-diagonalize the Hamiltonian, in the sense that it is closest to the identity on the space of configuration state functions (CSFs).[112–114] The CASVB structure of the CSF basis in Figure 2 was a significant motivation for my assertion of correspondence between the family of methine adapted SA-CASSCF solutions and the resonance theory. One might therefore expect that the angles parametrizing the 2 × 2 unitary transformations that diagonalize the minimum-norm and Platt Hamiltonians would be close to each other (in some reaonable metric[114]). I show that this is true in Figure 10, where the distribution of angle differences is plotted in a histogram. The distribution appears to have two components: a large spike at 0° (which includes all of the symmetric dyes in the set, plus a few asymmetric ones) and a broader peak centered near −5°. Since I parametrized the Platt Hamiltonian using the MS-MRPT2 corrected energies, the minimum-norm Hamiltonians used in the comparison also used these energies. The eigenvector information required to build the minimum-norm Hamiltonian[112] was taken from the MS-MRPT2-mixed SA-CASSCF eigenstates (as has been done in previous work[58]).

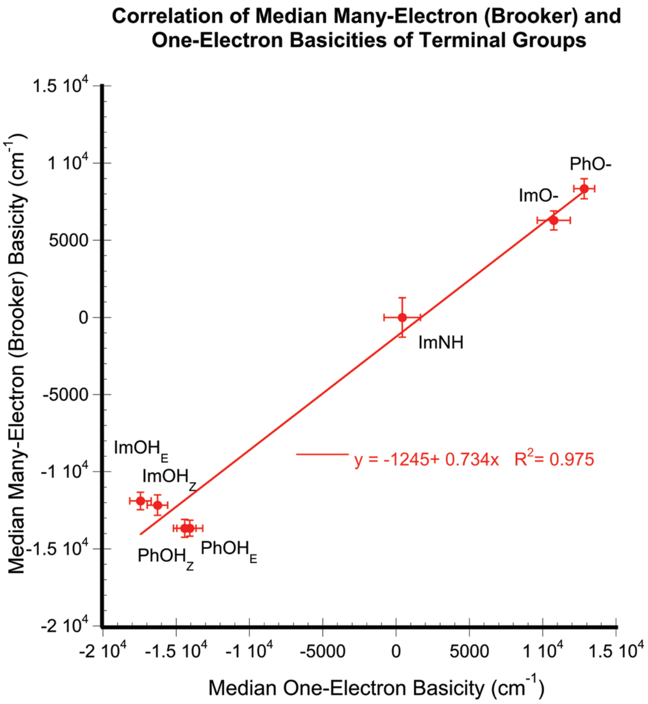


***Figure 9.*** Many-electron (Brooker) and one-electron basicity scales are strongly correlated. This shows that the Brooker (many-electron) basicity *does* represent a "basicity" in the usual (one-electron) sense. Brooker basicities were calculated using the calculated many-electron excitation energies as input, while the one-electron basicities were derived from the diagonal Fock matrix elements of Boys-localized active orbitals. The basicities shown are the medians of the distribution corresponding to each terminal group (see Figure 3). The error bars show the median absolute deviations of each distribution. I include a similar figure showing the distributions themselves in the Supporting Information. Deviation from the regression line indicates the importance of many-electron correlations to the basicity values.

The small magnitude of the difference angles in Figure 10 broadly supports my argument that the family of methine-adapted SA-CASSCF solutions corresponds to the resonance color theory. Interestingly, the correlation between the diagonal elements of the minimum-norm Hamiltonian and the one-electron basicity differences was significantly worse than the correlation of the many-electron and one-electron basicity differences over the dye set. It seems, therefore, that Platt's construction provides a higher fidelity mapping between one- and many-electron observables than does the minimum-norm block diagonalization.

## Discussion

I have shown that a modified version of the resonance color theory[14,16,18,19] can provide a systematic framework for understanding protonation-dependent color changes in an example set of monomethine oxonol dyes. I have done this in several steps. First, I have shown that there is a family of SA-CASSCF solutions which has the correct information structure for defining group basicities as was done in the theory, that this solution family has transferrability properties that mirror the Brooker basicity scale, and that the perturbed excitation energies calculated with this ansatz obey the Brooker deviation rule.[14,18] Second, I have shown that group-specific basicity indices can be extracted from the excitation energies of the set. When these are used to construct model Hamiltonians via Platt's construction,[14] the reconstructed excitations reproduce the input set to within the accuracy expected of the calculations themselves.[60,75,103–107] Third, I have demonstrated that these basicity indices are strongly correlated with an appropriately defined set of one-electron basicity indices, therefore showing that the representation is true to this core assertion of the resonance color theory. Finally, I have compared the Hamiltonians obtained via Platt's construction to an *ab initio* minimum-norm Hamiltonian in order to show that it is, in fact, an approximation to such a Hamiltonian.

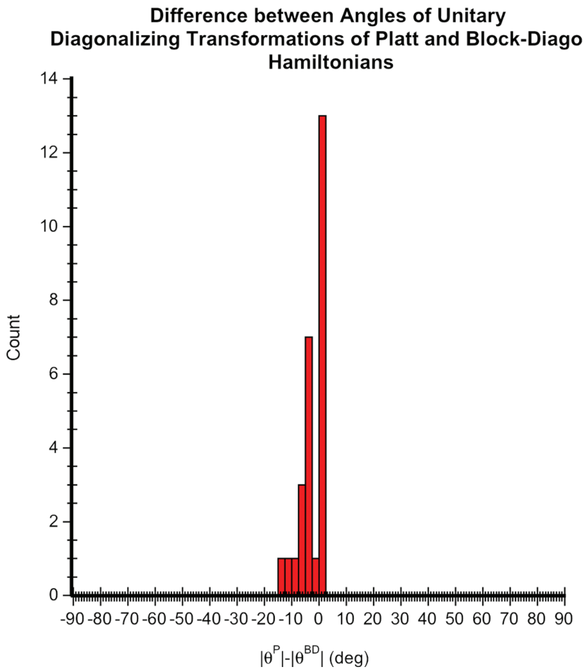


***Figure 10.*** The differences between angles that parametrize the transformations diagonalizing the Platt ($\theta^P$) and Block-Diagonalized ($\theta^{BD}$) effective Hamiltonians are clustered near zero. The Block Diagonalized Hamiltonian was generated by a minimum-norm block diagonalization of the Hamiltonian defined on the SA-CASSCF states in the Boys localized representation. The Platt Hamiltonian was constructed from the excitation energies of asymmetric and symmetric parent dyes as discussed in the text. The difference in the angle between the Hamiltonians is small, falling between 0° and −10° for nearly all dyes in the set. This shows that Platt's model Hamiltonian is a synthetic approximation to an *ab initio* effective Hamiltonian.

Although the results I report here were obtained for a relatively limited data set, my ongoing studies show that SA-CASSCF solutions of the form in Figure 2 can be readily obtained for a broad set of monomethine dyes. If this turns out to be generally true, then this provides a rigorous foundation for the resonance color theory in these systems. It immediately suggests a strategy for making quantitative, objective inferences using a conceptual framework that previously yielded only heuristic insights.

A formal validation of the resonance theory is important because the structure of the theory suggests many new strategies for modeling solvation and excited-state dynamics in complex environments. There has already been considerable work done in the field of nonequilbrium solvation, which, at least implicitly, invokes the resonance theory picture.[115–117] Similar models for nonlinear optical chromophores have also been proposed.[118–120] The "modular" nature of the relationship between color and group basicity in the theory is reminiscent of the modular nature of modern molecular dynamics force fields. This suggests that similar force fields for excited states may be achievable. This is a *very* interesting prospect, because there are a great many monomethine dyes which are already being used as probes of biomolecular structure and dynamics.[121–123] A better understanding of the excited states of these systems would add value to many results *already in the literature*, in addition to stimulating new work.

The data set I employed here includes the canonical protonation states of the GFP chromophore—specifically, the anion (PhO−, ImO−) and the phenolic neutral (PhOH, ImO−)—that are almost universally assigned to the "A" and "B" absorbance bands (Boxer's notation[98–100]) of GFPs.[10,124,125] It also includes the imidazol-3-ium "cationic" (PhOH, ImNH) and "zwitterionic" (PhO−, ImNH) forms which were once considered as candidates for assignment to these bands.[80,81] Our results suggest, as did previous ones, that this assignment would not be unreasonable if the absorption data alone were considered. Other data are contraindicative.[96,126] The set also includes two protonation states which, as far as I am aware, have not been assigned to the observed spectra: a dihydroxy cation (PhOH, ImOH) and an imidazolinol neutral form (PhO−, ImOH). The results I have obtained with these forms are interesting. First, I find that the (PhOH, ImOH) form, like the canonical (PhO−, ImO−) form, is *resonant* by Brooker's definition (small Brooker deviation) and is also quite red (446 nm). Though novel in this context, *this is not surprising* and could easily have been predicted on the basis of the *general* tendency of oxonol and hydroxyarylmethine dyes to develop deep color in acidic *and* alkaline solution.[2] The excitation energy of the (PhOH, PhOH) dye is bluer than the B bands of some GFP variants but is still comfortably in the range of wavelengths to which the label is applied.[9,127] Possibly, this deserves more attention. I have also found that the absorbances of the two neutral dyes (PhOH, ImO−) and (PhO−, ImOH) are very close (340 nm vs 342 nm) and would likely be *spectroscopically indistinguishable*. This suggests that the assignment of the A band in GFPs to a phenolic form should not be based on absorbance data alone. There is no immediate problem here: since their constituency is identical, the quantum chemical ground state energies of these forms can be directly compared. The ($PhOH_Z$, ImO−) form is lower in energy than the (PhO−, $ImOH_Z$) form by 29 kcal/mol, so there is no *a priori* case for reassignment.

Every fluorescent protein chromophore is synthesized inside the protein to which it is bound—*and with which it interacts*—for the entirety of its functional existence.[10] In the theory of open quantum systems, it is well-known that bipartitioning of a pure state system (having an idempotent density matrix) will lead to mixed state (nonidempotent) reduced density matrices for the subsystems, if the subsystems are entangled across the boundary.[128,129] If one naively assumes that the protein is described by a pure quantum state that is an eigenfunction of the proton number, then one is led, upon partitioning the chromophore from everything else, to consider that the most realistic assignment may include *ensembles* of protonation states. This possibility has already been suggested as an explanation for the spectra of positive-mode reversibly photoswitching GFPs.[130] The possibility that electronic and protonic states of the chromophore may be strongly coupled is consistent with the apparent ultrafast excited-state proton transfer speeds observed in some variants.[131] This topic is fit for discussion in a later publication.

A recent experimental study by Dong et al. of the solvatochromism of GFP chromophore models has suggested a high polarizability for the anion.[96] This is broadly consistent with our result that the anion is a resonant system near the methine limit.[3] In their study, Dong et al. fit the spectral solvatochromic data to a multivariate Kamlet−Taft equation[132] (linear solvatochromic energy relationship) and obtained interesting results. At pH values consistent with neutral (most likely (PhOH, ImO−)), the best-fit Kamlet−Taft coefficients were all negative (increasing solvent polarity, acidity, and (protic) basicity induced bathochromism). At pH values consistent with a cation (most likely (PhOH, ImNH)), all the coefficients were positive (hypsochromism). At pH values consistent with an anion, a bifurication was observed, so that increased polarity induced bathochromism while increases in acidity or basicity induced hypsochromism. The model compound used was not strictly the same as ours, differing by two methyl substituents to the imidazolinone. Methyl groups are weak electron donors, and the relevant active orbital in our model has amplitude at the points of their substitution, so one might expect the substitution to raise the orbital energy and affect a small basicity increase in the imidazolinone.

The Platt construction can, in principle, describe the bifurication of solvatochromic trends because it depends on two independent parameters.[14] For a particular case, the relevant question is how to partition solvation effects into the $E_I$'s and the $b$'s. I would argue that nonspecific interactions whose influence on the rings is anticorrelated (such as a dipole field component parallel to the long molecular axis) should manifest in the $b$'s alone, while ring-specific hydrogen bonding effects may be partly incorportated in the $E_I$ parameters (as I have done here, taking protonation as a limit of hydrogen-bond donation). A specific effect that is localized on one terminal group and independent of the other can be incorporated into the definition of a new symmetric parent structure. There is no easy way to "mirror" interactions that affect the rings in an anticorrelated way. In the case of a dipole field, the directionality of the field implies that "mirroring" the field would destroy its dipolar character. It is possible that higher-order field components would be preserved (for example certain quadrupolar fields), but most solvation models consider dipolar fields to be dominant contributors.[133] Interestingly, this partitioning into dipolar solvation effects and local (e.g., hydrogen-bonding) effects is exactly that suggested by the Kamlet−Taft analysis.[96] A bathochromic shift with increasing solvent polarity suggests a decrease in the absolute value of $b$(A,B), which would only be detected if the magnitude of $b$(A,B) was initially nonzero. It is possible, therefore, that the results of Dong et al. are indicative of electronic symmetry breaking in weakly polar solvents. It is interesting to note that the solvent with the highest dipolarity/polarizability parameter ($\pi^*$) studied by Dong et al. (water) reversed the trend toward increasing bathochromism. This may be indicative of crossover to a qualitatively distinct regime, but more data would be required to confirm or refute this idea. If protonation of the anion is considered as the extreme limit of hydrogen bond formation, then our results are consistent with hypsochromic shifts as the solvent hydrogen bond donation parameter is increased.

The resonance color theory predicts that the Brooker deviation should increase quadratically in the Brooker basicity difference. In this context, it is interesting to note the recent suggestion that a quadratic Stark effect is responsible for color tuning in red fluorescent proteins (RFPs).[134] RFPs are distinguished by an acylimine substitution to the imidazolinone ring. On the face of it, this suggests that the picture that emerges here for GFP chromophores may also be extended to chromophores from other subfamilies of the fluorescent proteins. A SA-CASSCF solution with the form of Figure 2 can also be obtained for an RFP chromophore model. In this case, there is one fragment orbital encompassing both the imidazolinone and acylimine moieties.

The resonance in the anionic dye (PhO−, ImO−) suggests high polarizability, as mentioned above. It is tempting to think that the prevalence of anionic chromophore forms in GFPs may have arisen *because* the resonant state allows greater flexibility for tuning the absorption. There are data supporting the idea that natural selection pressure drives emission color changes in these systems.[135−138] Not all GFP homologues are fluorescent, so it is conceivable that similar pressure drives the tuning of absorbance.

While this paper was in review, Martínez, Lamothe and I demonstrated that Brønsted acid/base chemistry and double bond photoisomerization chemistry are linked in GFP chromophores through the methine chemistry.[139] This suggests the interesting possibility that the structure−property relations I have described may also be applied to the photoisomerization reaction, which is considered to be a major decay channel in these systems.[124] The protonation-dependent basicity differences that I report here are *easily* larger than the $S_0$−$S_1$ energy gaps at favorable excited-state twisted configurations.[58,65,74,75,82,139] In fact, they are of similar magnitude to the *excitation energies themselves*. It is possible that this may be important to understanding why HBDI is nonfluorescent even in its crystalline solid phase, where large-amplitude twisting motions are unlikely to occur.[140]

## Conclusion

I have made the case that the resonance theory of Brooker[18] dyes provides a sound description for methine dye systems, if it is expressed in a revised model space that better reflects its information content. I have demonstrated the effectiveness of this approach by pointing out that there is a self-consistent model space for such dyes with the same form and demonstrated that, when the results obtainable with this solution are used to parametrize Platt's model Hamiltonian construction,[14] the set of excitations can be reconstructed without meaningful loss of accuracy. This provides a firmer theoretical foundation for the resonance theory and allows it to be used for quantitative, objective analyses where previously only heuristic insights could be obtained.

**Acknowledgment.** This work was supported by the Australian Research Council under Discovery Project DP0877875. Computations were done at the National Computational Infrastructure (NCI) Facility, Canberra, with time provided under Merit Allocation Scheme (MAS) project m03. I thank J. R. Reimers, N. S. Hush, and A. N. Tarnovsky for bringing Brooker's work to my attention; S. Boxer, S. Marder, S. Meech, T. Martínez, M. Prescott, M. Robb, M. Olivucci, G. Groenhof, L. Radom, T. Pullerits, T. Smith, M. Smith, S. Smith, and R. Jansen-Van Vuuren for helpful discussions; and R. McKenzie for research support, comments, and encouragement. Molecular graphics were generated using VMD.[141]

**Supporting Information Available:** Data pertaining to one-electron basicity extraction, Cartesian coordinates (Å), absolute SA-CASSCF and MS-MRPT2 energies (h), MS-MRPT2 mixing matrices, and SA-CASSCF natural and localized orbital graphics with state-averaged occupation numbers. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Lewis, G. N.; Calvin, M. *Chem. Rev.* **1939**, *25*, 273.

(2) Griffiths, J. *Colour and Constitution of Organic Molecules*; Academic Press Inc.: London, 1976; pp 2, 241; 248−250; 89−92; 258−259.

(3) Marder, S. R.; Gorman, C. B.; Meyers, F.; Perry, J. W.; Bourhill, G.; Brédas, J.-L.; Pierce, B. M. *Science* **1994**, *265*, 632.

(4) Wu, W.; Hua, J.; Jin, Y.; Zhan, W.; Tian, H. *Photochem. Photobiol. Sci.* **2008**, *7*, 63.

(5) Jain, V.; Rajbongshi, B. K.; Mallajosyula, A. T.; Bhattacharjya, G.; Iyer, S. S. K.; Ramanathan, G. *Sol. Energy Mater. Sol. Cells* **2008**, *92*, 1043.

(6) Puyol, M.; Encinas, C.; Rivera, L.; Miltsov, S. *Sens. Actuators* **2006**, *B115*, 287.

(7) Heilemann, M.; Margeat, E.; Kasper, R.; Sauer, M.; Tinnefeld, P. *J. Am. Chem. Soc.* **2005**, *127*, 3801.

(8) Weiss, P. S. *ACS Nano* **2008**, *2*, 1977.

(9) Tsien, R. Y. *Angew. Chem., Int. Ed.* **2009**, *48*, 5612.

(10) Remington, S. J. *Curr. Opin. Struct. Biol.* **2006**, *16*, 714.

(11) Shu, X.; Shaner, N. C.; Yarbrough, C. A.; Tsien, R. Y.; Remington, S. J. *Biochemistry* **2006**, *45*, 9639.

(12) Dahne, S. *Science* **1978**, *199*, 1163.

(13) Kuhn, H. *J. Chem. Phys.* **1949**, *17*, 1198.

(14) Platt, J. R. *J. Chem. Phys.* **1956**, *25*, 80.

(15) Dewar, M. *J. Chem. Soc.* **1950**, 2329.

(16) Moffitt, W. E. *Proc. Phys. Soc.* **1950**, *A63*, 700.

(17) Brooker, L. G. S.; Keyes, G. H.; Williams, W. W. *J. Am. Chem. Soc.* **1942**, *64*, 199.

(18) Brooker, L. G. S. *Rev. Mod. Phys.* **1942**, *14*, 275.

(19) Pauling, L. *Proc. Natl. Acad. Sci. U.S.A.* **1939**, *25*, 577.

(20) Fabian, J.; Hartmann, H. *Light Absorption of Organic Colorants*; Springer-Verlag: Heidelberg, Germany, 1980; pp 1−245.

(21) Berneth, H. In *Ullmann's Encyclopedia of Industrial Chemistry*; 7th ed.; John Wiley & Sons Inc.: New York, 2009.

(22) Platt, J. R. *J. Chem. Phys.* **1949**, *17*, 484.

(23) Herzfeld, K. F.; Sklar, A. L. *Rev. Mod. Phys.* **1942**, *14*, 0299.

(24) Reimers, J. R.; Hush, N. S. *Chem. Phys.* **1989**, *134*, 323.

(25) Feynman, R. P.; Leighton, R. B.; Sands, M. L. *Quantum Mechanics*; Addison-Wesley Publishing Company: Reading, MA, 1989; Vol. 3, pp 10−12.

(26) Mulliken, R. S. *J. Chem. Phys.* **1939**, *7*.

(27) Dewar, M. J. S.; Longuet-Higgins, H. C. *Proc. R. Soc. London* **1952**, *A214*, 482.

(28) Hoffmann, R.; Shaik, S.; Hiberty, P. C. *Acc. Chem. Res.* **2003**, *36*, 750.

(29) Shaik, S.; Hiberty, P. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Larter, R., Cundari, T. R., Eds.; John Wiley & Sons Inc.: New York, 2004; Vol. 20.

(30) Lipkin, H. J. *Lie Groups for Pedestrians*; Dover Publications Inc.: Mineola, NY, 2002; pp 16−17.

(31) Tishby, N. Z.; Levine, R. D. *Chem. Phys. Lett.* **1984**, *104*, 4.

(32) Shaik, S.; Shurki, A. *Angew. Chem., Int. Ed.* **1999**, *38*, 586.

(33) Truhlar, D. G. *J. Comput. Chem.* **2006**, *28*, 73.

(34) Hiberty, P. C. *THEOCHEM* **1998**, *451*, 237.

(35) Hiberty, P. C.; Shaik, S. *J. Comput. Chem.* **2007**, *28*, 137.

(36) Shaik, S. *New J. Chem.* **2007**, *31*, 1981.

(37) Hiberty, P. C.; Leforestier, C. *J. Am. Chem. Soc.* **1978**, *100*, 2012.

(38) Hirao, K.; Nakano, H.; Nakayama, K.; Dupuis, M. *J. Chem. Phys.* **1996**, *105*, 9227.

(39) Paldus, J.; Sarma, C. R. *J. Chem. Phys.* **1985**, *83*, 5135.

(40) Granucci, G.; Cassam-Chenaï, P.; Ellinger, Y. *J. Chem. Phys.* **1998**, *108*, 2538.

(41) Cassam-Chenaï, P.; Ellinger, Y.; Berthier, G. *Phys. Rev.* **1993**, *A48*, 2746.

(42) An exception worth noting is in the case where there are exactly two electrons in a closed active space. Then, the complete set of number-conserving operators spans a Lie algebra and is composed of one- and two- electron operators.

(43) Salem, L. *The Molecular Orbital Theory of Conjugated Systems*; W.A. Benjamin Inc.: New York, 1966; pp 36−43.

(44) Docken, K. K.; Hinze, J. *J. Chem. Phys.* **1972**, *57*, 4928.

(45) Stålring, J.; Bernhardsson, A.; Lindh, R. *Mol. Phys.* **2001**, *99*, 103.

(46) Roos, B. In *Adv. Chem. Phys.*; Lawley, K. P., Ed.; John Wiley & Sons Ltd.: New York, 1987; Vol. 69, p 399.

(47) Werner, H.-J.; Meyer, W. *J. Chem. Phys.* **1981**, *74*, 5794.

(48) Olsen, S.; McKenzie, R. H. *J. Chem. Phys.* **2009**, *131*, 234306.

(49) I have been able to obtain an analogous SA2-CAS(2,3) solution for the dication formed by removing two electrons from the GFP chromophore anion. The orbitals at convergence are visually indistinguishable from the four-electron case.

(50) Foster, J. M.; Boys, S. F. *Rev. Mod. Phys.* **1960**, *32*, 300.

(51) Levy, B.; Berthier, G. *Int. J. Quantum Chem.* **1968**, *2*, 307.

(52) Finley, J.; Malmqvist, P.; Roos, B.; Serrano-Andrés, L. *Chem. Phys. Lett.* **1998**, *288*, 299.

(53) Celani, P.; Werner, H.-J. *J. Chem. Phys.* **2003**, *119*, 5044.

(54) Celani, P.; Werner, H.-J. *J. Chem. Phys.* **2000**, *112*, 5546.

(55) Rajput, J.; Rahbek, D. B.; Andersen, L. H.; Rocha-Rinza, T.; Christiansen, O.; Bravaya, K. B.; Nemukhin, A. V.; Bochenkova, A. V.; Solntsev, K. M.; Dong, J.; Kowalik, J.; Tolbert, L. M.; Petersen, M. Å.; Nielsen, M. B. *Phys. Chem. Chem. Phys.* **2009**, *11*, 9996.

(56) Li, X.; Chung, L.; Mizuno, H.; Miyawaki, A.; Morokuma, K. *J. Phys. Chem. B* **2010**, *114*, 1114.

(57) Ma, Y.; Rohlfing, M.; Molteni, C. *J. Chem. Theory Comput.* **2010**, *5*, 257.

(58) Olsen, S.; McKenzie, R. H. *J. Chem. Phys.* **2009**, *130*, 184302.

(59) Polyakov, I.; Epifanovsky, E.; Grigorenko, B.; Krylov, A. I.; Nemukhin, A. *J. Chem. Theory Comput.* **2009**, *5*, 1907.

(60) Epifanovsky, E.; Polyakov, I.; Grigorenko, B.; Nemukhin, A.; Krylov, A. *J. Chem. Theory Comput.* **2009**, *5*, 1895.

(61) Filippi, C.; Zaccheddu, M.; Buda, F. *J. Chem. Theory Comput.* **2009**, *5*, 2074.

(62) Luin, S.; Voliani, V.; Lanza, G.; Bizzarri, R.; Amat, P.; Tozzini, V.; Serresi, M.; Beltram, F. *J. Am. Chem. Soc.* **2009**, *131*, 96.

(63) Bravaya, K. B.; Bochenkova, A. V.; Granovskii, A. A.; Nemukhin, A. V. *Russ. J. Phys. Chem.* **2008**, *B2*, 671.

(64) Vendrell, O.; Gelabert, R.; Moreno, M.; Lluch, J. M. *J. Chem. Theory Comput.* **2008**, *4*, 1138.

(65) Olsen, S.; Smith, S. *J. Am. Chem. Soc.* **2008**, *130*, 8677.

(66) Wang, S.; Smith, S. C. *Phys. Chem. Chem. Phys.* **2007**, *9*, 452.

(67) Camilloni, C.; Provasi, D.; Tiana, G.; Broglia, R. *J. Phys. Chem.* **2007**, *B111*, 10807.

(68) Zhang, L.; Xie, D.; Zeng, J. *J. Theory Comput. Chem.* **2006**, *5*, 375.

(69) Wang, S.; Smith, S. *J. Phys. Chem.* **2006**, *B110*, 5084.

(70) Zhang, R.; Nguyen, M. T.; Ceulemans, A. *Chem. Phys. Lett.* **2005**, *404*, 250.

(71) Xie, D.; Zeng, J. *J. Comput. Chem.* **2005**, *26*, 1487.

(72) Sinicropi, A.; Andruniow, T.; Ferre, N.; Basosi, R.; Olivucci, M. *J. Am. Chem. Soc.* **2005**, *127*, 11534.

(73) Vendrell, O.; Gelabert, R.; Moreno, M.; Lluch, J. M. *Chem. Phys. Lett.* **2004**, *396*, 202.

(74) Toniolo, A.; Olsen, S.; Manohar, L.; Martínez, T. J. *Faraday Disc.* **2004**, *127*, 149.

(75) Martin, M. E.; Negri, F.; Olivucci, M. *J. Am. Chem. Soc.* **2004**, *126*, 5452.

(76) Laino, T.; Nifosì, R.; Tozzini, V. *Chem. Phys.* **2004**, *298*, 17.

(77) Das, A.; Hasegawa, J.; Miyahara, T.; Ehara, M.; Nakatsuji, H. *J. Comput. Chem.* **2003**, *24*, 1421.

(78) Helms, V. *Curr. Opin. Struct. Biol.* **2002**, *12*, 169.

(79) Helms, V.; Winstead, C.; Langhoff, P. *THEOCHEM* **2000**, *506*, 179.

(80) Voityuk, A.; Michel-Beyerle, M.; Rösch, N. *Chem. Phys.* **1998**, *231*, 13.

(81) Voityuk, A.; Michel-Beyerle, M.; Rösch, N. *Chem. Phys. Lett.* **1997**, *272*, 162.

(82) Altoe, P.; Bernardi, F.; Garavelli, M.; Orlandi, G. *J. Am. Chem. Soc.* **2005**, *127*, 3952.

(83) Topol, I.; Collins, J.; Polyakov, I.; Grigorenko, B.; Nemukhin, A. *Biophys. Chem.* **2009**, *145*, 1.

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

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

(86) Bofill, J. M.; Pulay, P. *J. Chem. Phys.* **1989**, *90*, 3637.

(87) Pulay, P.; Hamilton, T. P. *J. Chem. Phys.* **1988**, *88*, 4926.

(88) Kutzelnigg, W. *Int. J. Quantum Chem.* **2009**, *109*, 3858.

(89) Canosa, N.; Rossignoli, R.; Plastino, A.; Miller, H. *Phys. Rev.* **1992**, *C45*, 1162.

(90) Malrieu, J.-P.; Heully, J.-L.; Zaitsevskii, A. *Theor. Chem. Acc.* **1995**, *90*, 167.

(91) MOLPRO, version 2009.1, a package of ab initio programs: Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M. and others, see http://www.molpro.net.

(92) There is one dye in the set that did violate the rule: the dye (PhOHZ/PhOHE). The excitation for this dye is very slightly below the mean wavelength of its parents, leading to an imaginary *b* value of *b*(PhOHZ/PhOHE) = 372i cm$^{-1}$. The magnitude of this number is very small, consistent with the (reasonable) expectation that the dye is near-resonant, and the orientation of the hydrogen atom has a small effect. In this case, I replaced the imaginary number with its magnitude, after verifying that this substitution led to a difference of less than 1 nm in the recalculated excitation wavelength.

(93) Forbes, M. W.; Jockusch, R. A. *J. Am. Chem. Soc.* **2009**, *131*, 17038.

(94) Andersen, L. H.; Lapierre, A.; Nielsen, S. B.; Nielsen, I. B.; Pedersen, S. U.; Pedersen, U. V.; Tomita, S. *Eur. Phys. J.* **2002**, *D20*, 597.

(95) Nielsen, S. B.; Lapierre, A.; Andersen, J. U.; Pedersen, U. V.; Tomita, S.; Andersen, L. H. *Phys. Rev. Lett.* **1997**, *87*, 228102.

(96) Dong, J.; Solntsev, K. M.; Tolbert, L. M. *J. Am. Chem. Soc.* **2006**, *128*, 12038.

(97) Webber, N. M.; Meech, S. R. *Photochem. Photobiol. Sci.* **2007**, *6*, 976.

(98) Shi, X.; Abbyad, P.; Shu, X.; Kallio, K.; Kanchanawong, P.; Childs, W.; Remington, S. J.; Boxer, S. G. *Biochemistry* **2007**, *46*, 12014.

(99) McAnaney, T.; Park, E.; Hanson, G.; Remington, S.; Boxer, S. *Biochemistry* **2002**, *41*, 15489.

(100) Chattoraj, M.; King, B.; Bublitz, G.; Boxer, S. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 8362.

(101) It is possible that the state-averaging procedure is removing contaminating continuum states from the active space, but this does not explain why they are not being re-introduced by the subsequent perturbation theory.

(102) Langley, R. *Practial Statistics Simply Explained*; Dover Publications Inc.: Mineola, NY, 1971; pp 222−230.

(103) Azizi, Z.; Roos, B.; Veryazov, V. *Phys. Chem. Chem. Phys.* **2006**, *8*, 2727.

(104) Serrano-Andrés, L.; Merchán, M. *THEOCHEM* **2005**, *729*, 99.

(105) Andersen, L. H.; Bochenkova, A. V. *Eur. Phys. J.* **2009**, *D51*, 5.

(106) van Faassen, M.; de Boeij, P. L. *J. Chem. Phys.* **2004**, *120*, 11967.

(107) Schreiber, M.; Silva-Junior, M. R.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2008**, *128*, 134110.

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

(109) Jacquemin, D.; Wathelet, V.; Perpète, E. A.; Adamo, C. *J. Chem. Theory Comput.* **2009**, *5*, 2420.

(110) In particular cases, accuracies exceeding these *a priori* expectations are achiveable. See, for example: Schreiber, M.; Buss, V.; Fulscher, M. P. *Phys. Chem. Chem. Phys.* **2001**, *3*, 3906.

(111) Breuer, H.-P.; Petruccione, F. *The Theory of Open Quantum Systems*; Oxford University Press: Oxford, U.K., 2002; pp 5−10.

(112) Cederbaum, L. S.; Schirmer, J.; Meyer, H.-J. *J. Phys. (Paris)* **1989**, *A22*, 2427.

(113) Goldstein, J. A.; Levy, M. *Am. Math. Monthly* **1991**, *98*, 710.

(114) Dupré, M.; Goldstein, J.; Levy, M. *J. Chem. Phys.* **1980**, *72*, 780.

(115) Bianco, R.; Timoneda, J.; Hynes, J. *J. Phys. Chem.* **1994**, *98*, 12103.

(116) Kim, H.; Hynes, J. T. *J. Chem. Phys.* **1990**, *93*, 5194.

(117) Kim, H.; Hynes, J. T. *J. Chem. Phys.* **1990**, *93*, 5211.

(118) Thompson, W.; Blanchard-Desce, M.; Alain, V.; Muller, J.; Fort, A.; Barzoukas, M.; Hynes, J. T. *J. Phys. Chem.* **1999**, *A103*, 3766.

(119) Thompson, W.; Blanchard-Desce, M.; Hynes, J. T. *J. Phys. Chem.* **1998**, *A102*, 7712.

(120) Lu, D.; Chen, G.; Perry, J.; Goddard, W. A., III. *J. Am. Chem. Soc.* **1994**, *116*, 10679.

(121) Babendure, J.; Adams, S.; Tsien, R. *J. Am. Chem. Soc.* **2003**, *125*, 14716.

(122) Gonçalves, M. *Chem. Rev* **2009**, *109*, 190.

(123) Ozhalici-Unal, H.; Pow, C. L.; Marks, S. A.; Jesper, L. D.; Silva, G. L.; Shank, N. I.; Jones, E. W.; Burnette, J. M.; Berget, P. B.; Armitage, B. A. *J. Am. Chem. Soc.* **2008**, *130*, 12620.

(124) Meech, S. R. *Chem. Soc. Rev.* **2009**, *38*, 2922.

(125) Tsien, R. Y. *Annu. Rev. Biochem.* **1998**, *67*, 509.

(126) Bell, A. F.; He, X.; Wachter, R.; Tonge, P. J. *Biochemistry* **2000**, *39*, 4423.

(127) Pakhomov, A. A.; Martynov, V. *Chem. Biol.* **2008**, *15*, 755.

(128) Jaynes, E. T. *Phys. Rev.* **1957**, *108*, 171.

(129) Bengtsson, I.; Życzkowski, K. *The Geometry of Quantum States*; Cambridge University Press: Cambridge, U. K., 2006; pp 333−369.

(130) Andresen, M.; Stiel, A. C.; Fölling, J.; Wenzel, D.; Schönle, A.; Egner, A.; Eggeling, C.; Hell, S. W.; Jakobs, S. *Nat. Biotechnol.* **2008**, *26*, 1035.

(131) Kondo, M.; Heisler, I. A.; Stoner-Ma, D.; Tonge, P. J.; Meech, S. R. *J. Am. Chem. Soc.* **2009**, *132*, 1452.

(132) Kamlet, M. J.; Abboud, J. L.; Abraham, M. H.; Taft, R. W. *J. Org. Chem.* **1983**, *48*, 2877.

(133) Kamlet, M. J.; Abboud, J. L.; Taft, R. W. *J. Am. Chem. Soc.* **1977**, *99*, 6027.

(134) Drobizhev, M.; Tillo, S.; Makarov, N. S.; Hughes, T. E.; Rebane, A. *J. Phys. Chem.* **2009**, *B113*, 12860.

(135) Gruber, D. F.; Desalle, R.; Lienau, E. K.; Tchernov, D.; Pieribone, V. A.; Kao, H.-T. *Mol. Bio. Evol.* **2009**, *26*, 2841.

(136) Labas, Y. A.; Gurskaya, N. G.; Yanushevich, Y. G.; Fradkov, A. F.; Lukyanov, K. A.; Lukyanov, S. A.; Matz, M. V. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 4256.

(137) Field, S. F.; Bulina, M. Y.; Kelmanson, I. V.; Bielawski, J. P.; Matz, M. V. *J. Mol. Evol.* **2006**, *62*, 332.

(138) Ugalde, J.; Chang, B.; Matz, M. *Science* **2004**, *305*, 1433.

(139) Olsen, S.; Lamothe, K.; Martínez, T. J. *J. Am. Chem. Soc.* **2010**, *132*, 1192.

(140) Dong, J.; Solntsev, K. M.; Tolbert, L. M. *J. Am. Chem. Soc.* **2009**, *131*, 662.

(141) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graph.* **1996**, *14*, 33.

CT100001B

JCTC Journal of Chemical Theory and Computation

# Density Functional Calculations of E2 and $S_N2$ Reactions: Effects of the Choice of Density Functional, Basis Set, and Self-Consistent Iterations

Yan Zhao[†] and Donald G. Truhlar[*,‡]

*Commercial Print Engine Lab, Hewlett-Packard Laboratories, Hewlett-Packard Company, 1501 Page Mill Road, Palo Alto, California 94304 and Department of Chemistry and Supercomputing Institute, University of Minnesota, 207 Pleasant Street S.E., Minneapolis, Minnesota 55455-0431*

Received January 10, 2010

**Abstract:** We have computed stationary points on the potential energy surface for the *anti*-E2, *syn*-E2, and $S_N2$ pathways of the reactions of $F^-$ and $Cl^-$ with $CH_3CH_2F$ and $CH_3CH_2Cl$ with fully self-consistent fields and Gaussian basis functions. We find large differences from previously reported [Bento, A. P.; Solà, M.; Bickelhaupt, F. M. *J. Chem. Theory Comput.* **2008**, *4*, 929] calculations with Slater-type orbitals. We revise the findings of the previous study; in particular, we find average absolute errors in kcal/mol compared to benchmark calculations of 20 stationary point energies (6 saddle points and 14 minima) of 0.9 for M06-2X, 1.2 for M08-SO, 1.4 for M06-HF, 2.0 for M06, 2.3 for B3LYP, 2.5 for OLYP, 2.7 for M06-L, and 3.5 kcal/mol for TPSS. We also compare the predictions of various density functionals for the partial atomic charges at the transition states.

## 1. Introduction

Density functional theory[1] has become a powerful tool for predicting and understanding trends in chemical reactivity, and considerable time has been expended in validating its predictions.[2] A recent article[3] in this journal presented calculations for bimolecular elimination (E2) and bimolecular nucleophilic substitution ($S_N2$) reactions with a variety of density functional approximations (DFAs) and compared the results to benchmark calculations. The article found large errors in the predictions of most DFAs in the literature; it found a mean unsigned error for M06-2X of 2.3 kcal/mol, and it found that the M06-L density functional, although being among the best functionals in terms of mean unsigned errors (MUEs) in barrier heights, incorrectly predicted that the $S_N2$ saddle point for the reaction of $F^-$ with $CH_3CH_2F$ is lower in energy than that of the *anti*-E2 saddle point. In order to study this, we began to repeat some of their calculations with a different computer program, and we found surprisingly large differences.

In this article we present our results, compare them to their results, and update the conclusions about the accuracy of various density functionals. We also provide tests of the post-SCF approximation and the density fitting procedure, we test two additional density functionals that were not included in ref 3, and we calculate the charge distributions at the transition states.

## 2. Methods

We explored several methods where each method is a combination of a DFA and a basis set. The DFAs[4−13] considered here are summarized in Table 1.

We considered several Gaussian basis sets,[14−20] which are specified in Table 2, which gives convenient abbreviations for use in the rest of the article. The final row of Table 2 is the Slater-type basis used in the calculations of ref 3, which were all performed with the ADF program.[21] Our calculations in Tables 3 and 4 were performed with the Gaussian 03[22] and MN-GFM programs.[23]

[*] Corresponding author. Telephone: 612-624-7555. Fax: 612-624-9390. E-mail: truhlar@umn.edu.

[†] Hewlett-Packard Company.

[‡] University of Minnesota.

10.1021/ct100082z
Published on Web 03/17/2010

***Table 1.*** Density Functional Approximations[a]

| name | reference | type | $X^b$ |
|---|---|---|---|
| OLYP | 6, 9 | GGA | 0 |
| B3LYP | 5−9 | hybrid GGA | 20 |
| TPSS | 10 | meta-GGA | 0 |
| M06-L | 4 | meta-GGA | 0 |
| M06-HF | 11 | hybrid meta[c] | 100 |
| M06 | 12 | hybrid meta | 27 |
| M06-2X | 12 | hybrid meta | 54 |
| M08-HX | 13 | hybrid meta | 52.23 |
| M08-SO | 13 | hybrid meta | 56.79 |

[a] We sometimes use the conventional language in which an approximation to the unknown exact exchange−correlation functional is just called a density functional. [b] *X* denotes the percentage of Hartree−Fock exchange. [c] hybrid meta is short notation for hybrid meta-GGA.

***Table 2.*** Basis Sets

| basis | reference | type[a] (quality) | abbreviation |
|---|---|---|---|
| 6-311+G(2df,2p) | 14, 15 | triple-ζ + diffuse | 311+ |
| MG3S[b] | 16, 17 | triple-ζ + diffuse | MG3S |
| aug-cc-pV(T+d)Z | 18 | triple-ζ + diffuse | aTZ |
| aug-cc-pV(Q+d)Z | 18, 19 | quadruple-ζ + diffuse | aQZ |
| TZ2P | 20 | Slater triple-ζ + diffuse | V |

[a] "triple-ζ quality" is short notation for polarized triple-ζ. [b] MG3S is the same as 6-311+G(2df,2p) for H, C, and F but is improved for Cl. See ref 17.

For all calculations, we used the default fine grid in *Gaussian 03*, which has 75 radial shells and 302 angular points per shell. We have checked the M06-2X/MG3S calculations with the ultrafine grid (99 radial shells and 590 angular points per shell), and we found that the relative energies obtained with the default fine grids agree within 0.05 kcal/mol of those obtained by using ultrafine grids.

The *Gaussian 03* program can carry out Kohn−Sham density functional calculations with or without density fitting,[24] and we used both methods, as will be specified in the tables. In principle, density fitting is not an approximation, if the auxiliary basis set used for fitting is converged, but in practice one should check for basis set incompleteness or use well validated large auxiliary basis sets.

## 3. Results

Tables 3 and 4 show the present results for the $F^- + CH_3CH_2F$ and $Cl^- + CH_3CH_2Cl$ reactions, respectively, and Tables 5 and 6 present selected results from Bento et al.,[3] arranged the same way for ease of comparison. As in ref 3, we consider the *syn*- and *anti*-E2 reactions and the $S_N2$ reaction. All calculations in Tables 3−6 are based on the OLYP/TZ2P geometries of Bento et al.[3] in order to remove the choice of geometry from the comparison of calculations. In the tables, the symbol "A@B" is used to denote a post-SCF calculation, that is, a calculation in which the final energy is evaluated by method A, but the density is calculated by method B.

All values in the tables are energies relative to the energy of the separated reactants. PC denotes a precursor complex (local minimum of the potential energy occurring at an earlier stage of the reaction coordinate than the saddle point), SP denotes the saddle point, SC denotes a successor complex (local minimum of the potential energy occurring at a later stage of the reaction coordinate than the saddle point), and P denotes the product.

## 4. Discussion

**4.1. Comparison of Calculations.** In addition to density functional calculations, Tables 5 and 6 show benchmark calculations carried out by the coupled cluster method with single and double excitations and a quasiperturbative treatment of connected triple excitations[25] (CCSD(T)), extrapolated[3] to the complete basis set (CBS) limit. These benchmark calculations are used to compute the mean unsigned error (MUE) for each row of Tables 3−6, and these MUEs are given in the last column of each of these tables. The first observation one makes is that the errors are much bigger in Tables 5 and 6, based on the Slater-type orbitals (STOs),[3] than in Tables 3 and 4, based on the Gaussian-type orbital calculations (GTO calculations) presented here.

The striking differences in the results must be attributed to the differences in the calculations: the different basis sets, the use of post-SCF calculations in some of the STO calculations, and the use of frozen core approximation in some of the STO calculations.

The first possible reason for the differences is that we use Gaussian basis sets and ADF uses Slater-type orbitals. However, the first three rows of Tables 3 and 4 show that we get similar results with two different triple-ζ quality and one quadruple-ζ quality basis set, so it does not appear that basis set incompleteness is significant in our GTO calculations. Personal communications from Bickelhaupt and van Gisbergen and their co-workers show that the largest systematic source of discrepancy between the results of ref 3 and the present results is the choice of STO basis set; in particular, a large part of the discrepancy can be removed if one uses larger STO basis sets containing diffuse functions.[26,27] It is not our goal to trace down every possible contributor to the deviations between the results of ref 3 and the present results but rather to focus on the revised conclusions about the performances of various density functionals, to present results for two density functionals (M08-HX and M08-SO) that were not included in ref 3, and to add a caution about the post-SCF approximation.

ADF can evaluate the energy for the density functionals of interest from a charge density obtained with a simpler density functional; this is called a post-SCF treatment, and it is provided as a convenience to user who wish to get a quick impression of the effect of changing the density functional. In the calculations of Bento et al., all charge densities were obtained with the OLYP functional. Thus all their results except OLYP involve the post-SCF approximation. In contrast, except for a test that we will now report, our calculations were all performed by achieving a full self-consistent field for each function employed. Recently Liao et al.[28] indicated that the post-SCF approximation can produce large errors in calculating noncovalent binding energies; in contrast, the authors of ref 3 state that in previous work[29] the post-SCF approximation has "been extensively tested and. .. shown to introduce an error in the computed energies of only a few tenths of a kcal/mol," but they did

***Table 3.*** Results with Gaussian Basis Sets for the Reactions of $F^-$ with $CH_3CH_2F$

| | *anti*-E2 | | | | *syn*-E2 | | | | $S_N2$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| method | PC | SP | SC | P | PC | SP | SC | P | PC | SP | MUE |
| M06-2X/MG3S | −15.99 | −1.15 | −5.49 | 19.16 | −12.07 | 3.85 | −35.12 | −29.47 | −15.99 | 3.02 | 1.34 |
| M06-2X/aTZ | −15.10 | −0.42 | −4.82 | 18.89 | −11.18 | 4.92 | −34.45 | −28.95 | −15.10 | 3.65 | 1.16 |
| M06-2X/aQZ | −15.05 | −0.25 | −4.74 | 18.72 | −11.20 | 5.17 | −33.99 | −28.54 | −15.05 | 4.07 | 1.20 |
| M06-2X@OLYP/aTZ | −13.60 | 2.33 | −3.63 | 18.54 | −9.89 | 6.24 | −34.24 | −28.74 | −13.60 | 6.26 | 2.07 |
| M06-2X/311+ | −15.99 | −1.15 | −5.49 | 19.16 | −12.07 | 3.85 | −35.12 | −29.47 | −15.99 | 3.02 | 1.34 |
| OLYP/aTZ | −12.51 | −0.74 | −4.55 | 12.97 | −8.25 | 4.17 | −31.65 | −28.15 | −12.51 | 4.16 | 2.23 |
| OLYP/aQZ | −12.21 | −0.39 | −4.25 | 12.94 | −7.98 | 4.52 | −31.19 | −27.80 | −12.21 | 4.58 | 2.47 |
| OLYP/311+ | −14.63 | −2.63 | −6.39 | 13.54 | −10.16 | 1.69 | −33.45 | −29.37 | −14.63 | 1.35 | 1.45 |
| B3LYP/311+ | −15.98 | −2.65 | −6.78 | 16.38 | −11.59 | 2.31 | −34.82 | −29.98 | −15.98 | −1.08 | 1.58 |
| TPSS/311+ | −17.54 | −1.75 | −4.78 | 20.24 | −12.73 | 0.65 | −32.06 | −26.98 | −17.54 | −6.26 | 3.40 |
| M06/311+ | −16.25 | −1.76 | −5.84 | 18.49 | −12.22 | 2.86 | −31.70 | −26.58 | −16.25 | −1.02 | 2.14 |
| M06-L/311+ | −17.07 | −1.95 | −5.41 | 19.32 | −12.66 | 1.69 | −31.04 | −25.92 | −17.07 | −4.05 | 3.05 |
| M06-L/311+[a] | −17.06 | −1.95 | −5.41 | 19.32 | −12.65 | 1.70 | −31.04 | −25.93 | −17.06 | −4.03 | 3.04 |
| M08-HX/311+ | −16.56 | −1.96 | −6.47 | 18.58 | −12.64 | 3.90 | −34.38 | −28.67 | −16.56 | 1.26 | 1.44 |
| M08-SO/311+ | −15.87 | −2.44 | −6.13 | 18.21 | −12.22 | 4.83 | −33.35 | −27.77 | −15.87 | 0.89 | 1.40 |
| M06-HF/311+ | −15.80 | −0.79 | −4.98 | 19.75 | −11.76 | 4.48 | −38.51 | −32.77 | −15.80 | 2.92 | 1.56 |

[a] This calculation was carried out with density fitting; all other calculations in this table were carried out without this additional approximation.

***Table 4.*** Results with Gaussian Basis Sets for the Reactions of $Cl^-$ with $CH_3CH_2Cl$

| | *anti*-E2 | | | | *syn*-E2 | | | | $S_N2$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| method | PC | SP | SC | P | PC | SP | SC | P | PC | SP | MUE |
| M06-2X/MG3S | −11.15 | 17.64 | 11.56 | 23.26 | −11.15 | 29.76 | −4.54 | −1.69 | −11.15 | 5.79 | 0.54 |
| M06-2X/aTZ | −11.06 | 17.71 | 11.56 | 23.32 | −11.06 | 29.76 | −4.47 | −1.60 | −11.06 | 5.95 | 0.53 |
| M06-2X/aQZ | −10.91 | 18.10 | 11.83 | 23.39 | −10.91 | 30.24 | −4.08 | −1.26 | −10.91 | 6.49 | 0.61 |
| M06-2X@OLYP/aTZ | −10.38 | 20.08 | 12.10 | 22.94 | −10.38 | 31.80 | −4.02 | −1.17 | −10.38 | 7.81 | 1.10 |
| M06-2X/311+ | −11.34 | 17.51 | 11.71 | 23.72 | −11.34 | 29.61 | −4.83 | −1.80 | −11.34 | 5.63 | 0.69 |
| OLYP/aTZ | −8.03 | 13.48 | 9.42 | 16.65 | −8.03 | 22.62 | −8.32 | −6.46 | −8.03 | 6.87 | 3.76 |
| OLYP/aQZ | −7.91 | 13.79 | 9.74 | 16.88 | −7.91 | 22.93 | −7.97 | −6.11 | −7.91 | 7.24 | 3.64 |
| OLYP/311+ | −8.18 | 13.12 | 9.52 | 17.20 | −8.18 | 22.32 | −8.73 | −6.84 | −8.18 | 6.45 | 3.75 |
| B3LYP/311+ | −9.89 | 13.19 | 8.65 | 19.08 | −9.89 | 23.64 | −7.81 | −5.59 | −9.89 | 2.50 | 3.05 |
| TPSS/311+ | −10.42 | 10.62 | 10.53 | 21.95 | −10.42 | 19.61 | −7.56 | −4.99 | −10.42 | −1.73 | 3.56 |
| M06/311+ | −11.37 | 15.97 | 12.08 | 24.21 | −11.37 | 25.59 | −4.01 | −0.87 | −11.37 | 2.25 | 1.77 |
| M06-L/311+ | −11.11 | 13.33 | 11.37 | 23.39 | −11.11 | 22.41 | −5.07 | −2.01 | −11.11 | 0.05 | 2.29 |
| M06-L/311+[a] | −11.11 | 13.32 | 11.35 | 23.37 | −11.11 | 22.40 | −5.10 | −2.04 | −11.11 | 0.03 | 2.29 |
| M08-HX/311+ | −11.64 | 18.21 | 10.77 | 23.65 | −11.64 | 31.36 | −2.89 | 0.30 | −11.64 | 6.37 | 0.89 |
| M08-SO/311+ | −11.64 | 17.21 | 10.36 | 23.37 | −11.64 | 29.91 | −3.33 | −0.11 | −11.64 | 4.37 | 0.98 |
| M06-HF/311+ | −11.25 | 19.31 | 12.43 | 24.16 | −11.25 | 34.28 | −3.59 | −0.39 | −11.25 | 6.40 | 1.26 |

[a] This calculation was carried out with density fitting; all other calculations in this table were carried out without this additional approximation.

***Table 5.*** Results with Slater-type Basis Sets for the Reactions of $F^-$ with $CH_3CH_2F$

| | *anti*-E2 | | | | *syn*-E2 | | | | $S_N2$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| method | PC | SP | SC | P | PC | SP | SC | P | PC | SP | MUE |
| CCSD(T)/CBS | −14.89 | −1.27 | −6.35 | 15.77 | −11.00 | 5.68 | −37.39 | −28.60 | −14.89 | 2.20 | |
| B3LYP@OLYP/V | −19.30 | −5.38 | −10.66 | 15.90 | −14.50 | −2.00 | −40.32 | −35.34 | −19.30 | −4.01 | 4.44 |
| M06@OLYP/V | −18.21 | −2.21 | −7.47 | 17.88 | −13.96 | 1.14 | −35.19 | −30.59 | −18.21 | −0.35 | 2.51 |
| M06-2X@OLYP/V | −15.67 | 1.49 | −5.62 | 18.37 | −11.47 | 4.03 | −37.77 | −32.90 | −15.67 | 5.82 | 1.81 |
| TPSS@OLYP/V | −21.38 | −5.26 | −8.94 | 19.81 | −16.28 | −4.16 | −37.83 | −32.52 | −21.38 | −10.03 | 5.53 |
| M06-L@OLYP/V | −20.04 | −1.23 | −5.44 | 20.54 | −15.33 | 1.68 | −32.57 | −27.78 | −20.04 | −2.95 | 3.51 |
| OLYP/V | −20.01 | −7.95 | −12.49 | −12.85 | −15.20 | −4.93 | −41.40 | −36.41 | −20.01 | −4.16 | 8.47 |

not report tests of these approximations for the reactions under consideration here. We, therefore, tested the post-SCF method for these reactions by carrying out calculations for M06-2X with the OLYP density. These results are given in row four of Tables 3 and 4. The results differ from straight M06-2X calculations with the same basis set by 0.2−2.8 kcal/mol, with an average (over 20 cases) of 1.2 kcal/mol. Clearly the errors are larger than the previous work led the authors of ref 3 to believe, but the deviations of the present results from those of ref 3 are even larger than this, so this does not contradict the conclusion[26,27] that the main reason for the inaccuracy of ref 3 is the choice of basis set.

Our OLYP calculations are also important for another reason. Since OLYP is the only method for which the results in ref 3 are full SCF results rather than post-SCF results, this comparison directly tests whether issues other than the post-SCF approximation are indeed important. The results show that they are. Comparing our results with the largest basis set (aQZ) to the results from ref 3 shows absolute deviations of 0.1−10.2 kcal/mol, with an average (over 20

***Table 6.*** Results with Slater-type Basis Sets for the Reactions of $Cl^-$ with $CH_3CH_2Cl$

| | *anti*-E2 | | | | *syn*-E2 | | | | $S_N2$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| method | PC | SP | SC | P | PC | SP | SC | P | PC | SP | MUE |
| CCSD(T)/CBS | −11.07 | 18.18 | 9.77 | 22.16 | −11.07 | 30.92 | −4.85 | −1.42 | −11.07 | 5.81 | |
| B3LYP@OLYP/V | −10.60 | 11.00 | 7.03 | 17.83 | −10.60 | 21.22 | −10.78 | −8.53 | −10.60 | 0.92 | 4.33 |
| M06@OLYP/V | −12.68 | 17.33 | 6.17 | 22.96 | −12.68 | 23.67 | −6.49 | −2.92 | −12.68 | 3.36 | 2.29 |
| M06-2X@OLYP/V | −12.49 | 10.65 | 14.98 | 22.49 | −12.49 | 30.29 | −5.85 | −4.74 | −12.49 | 10.73 | 2.72 |
| TPSS@OLYP/V | −10.99 | 9.34 | 8.95 | 20.88 | −10.99 | 17.58 | −9.89 | −7.01 | −10.99 | −3.22 | 4.42 |
| M06-L@OLYP/V | −14.02 | 12.92 | 8.99 | 25.92 | −14.02 | 22.77 | −4.73 | −2.25 | −14.02 | 2.63 | 3.09 |
| OLYP/V | −9.66 | 10.68 | 7.45 | 16.33 | −9.66 | 19.58 | −11.81 | −9.28 | −9.66 | 4.04 | 4.78 |

cases) of 5.0 kcal/mol. We note for completeness that the OLYP calculations of ref 3 were carried with a frozen-core approximation, whereas all other calculations in ref 3 were all-electron calculations; our calculations are all of the latter type.

ADF evaluates the matrix elements of the exchange−correlation potential and the Coulomb potential using an auxiliary function representation of the electronic density; this is called density fitting. Gaussian can carry out calculations with or without this additional approximation. When density fitting is employed, it is important to know the adequacy of auxiliary basis sets used for density fitting, Tables 3 and 4 show M06-L calculations with and without this approximation, using the *Gaussian 03* default choice for the auxiliary basis, and the agreement is quite good, with an average deviation of only 0.01 kcal/mol. This is consistent with conclusions we drew in tests of this method on other reactions.[2f] We did not test the auxiliary basis sets used in ADF, but the sets of auxiliary functions in ADF are rather extensive, and tests of this feature by the developers of ADF show that density fitting errors are quite small (smaller than basis set effects).[21]

**4.2. Revised Conclusions about the Accuracy of Density Functional Approximations.** The mean unsigned error in the present M06-2X results is 0.9 kcal/mol, as compared to 2.3 kcal/mol in the calculations of ref 3. The average error in the M06-L relative energetics (average over 20 cases) is 2.7 kcal/mol, which is lower than the 3.3 kcal/mol average error found in ref 3. We conclude that the bigger-than-expected errors reported in ref 3 appear to be artifacts of the calculations, especially the basis sets, and not deficiencies of the density functionals applied.

We still find that the M06-L density functional predicts an anomalously low barrier height for the $F^- + CH_3CH_2F$ $S_N2$ reaction, and in fact, it also predicts a significantly too low barrier for the $Cl^- + CH_3CH_2Cl$ $S_N2$ reaction. Since functionals with no Hartree−Fock exchange sometimes overestimate the amount of charge transfer,[31] one possible reason for the large error in M06-L in the $F^-$ case is an overestimate of the amount of charge transfer. To examine this possibility, we computed the partial atomic charges at transition states by Hirshfeld population analysis.[32,33] The partial atomic charges on the halogen atom are in Table 7, which shows reasonably similar partial atomic charges for various density functionals. Furthermore, the $S_N2$ reaction has neither the largest nor the smallest partial charges on the halogens. Thus we cannot attribute the poor performance of $S_N2$ to spurious charge transfer. M06-L is a meta-GGA. The only other meta-GGA in this paper is TPSS; TPSS has an even larger error than M06-L for this difficult case.

***Table 7.*** Hirshfeld Partial Atomic Charges on F and Cl at the *anti*-E2 and $S_N2$ Transition States[a]

| | *anti*-E2 | | | | $S_N2$[b] | |
|---|---|---|---|---|---|---|
| method | leaving F | incoming F | leaving Cl | incoming Cl | F | Cl |
| M06-L | −0.74 | −0.29 | −0.73 | −0.39 | −0.51 | −0.58 |
| M06-2X | −0.80 | −0.29 | −0.77 | −0.38 | −0.53 | −0.59 |
| M06-HF | −0.82 | −0.29 | −0.78 | −0.36 | −0.55 | −0.59 |
| OLYP | −0.72 | −0.28 | −0.70 | −0.38 | −0.49 | −0.55 |
| TPSS | −0.73 | −0.28 | −0.71 | −0.38 | −0.49 | −0.56 |
| HF | −0.84 | −0.31 | −0.81 | −0.39 | −0.60 | −0.65 |

[a] All calculations in this table employed the 311+ basis set. [b] The incoming and leaving halogen atoms have the same partial charges for the $S_N2$ reactions.

The average error in the M06-L relative energetics (average over 20 cases) is 2.7 kcal/mol, which is lower than our calculated mean errors for other functionals with no Hartree−Fock exchange: 3.1 and 3.5 kcal/mol for OLYP and TPSS, respectively. Adding Hartree−Fock exchange can reduce the error to 0.9 (M06-2X), 1.2 (M08-HX and M08-SO), 1.4 (M06-HF), or 2.0 kcal/mol (M06), whereas the popular B3LYP functional has a mean unsigned error of 2.3 kcal/mol.

Excluding one saddle point (that for the $S_N2$ reaction of $F^-$ with $CH_3CH_2F$), reduces the mean absolute errors for of M06-2X and M06-L from respectively 0.89 and 2.67 kcal/mol for 20 cases to 0.85 and 2.48 kcal/mol for 19 cases.

## 5. Concluding Remarks

The density functional results obtained here agree much better than those in ref 3 with benchmark calculations. The final trends in the errors for barrier heights in the present paper are not significantly out of line with what might have been expected on the basis of previous tests,[2e,2f] although the errors reported in ref 3 are larger than would have been expected for the Minnesota functionals (which was the motivation for our opening the problem for re-examination). The errors in the calculations with the more accurate density functionals could probably be reduced by using consistently optimized geometries rather than OLYP geometries, but that is not the goal of the present article, which is instead designed to reveal differences in the calculated results due to algorithmic choices in ref 3.

**Acknowledgment.** The authors are very grateful to Stan van Gisbergen, Erik van Lenthe, F. Matthias Bickelhaupt, Marcel Swart, and Miquel Solà for assistance in tracing down the discrepancies between the present calculations and those of ref 3. The research reported here was supported in part by the Air Force Office of Scientific Research.

## References

(1) Kohn, W.; Becke, A. D.; Parr, R. G. *J. Phys. Chem.* **1996**, *100*, 12974.

(2) See, e.g., (a) Baker, J.; Muir, M.; Andzelm, J.; Scheiner, A. *ACS Symp. Ser.* **1996**, *629*, 342. (b) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K.; Pople, J. A. *J. Chem. Phys.* **2000**, *112*, 7374. (c) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **2005**, *123*, 124107. (d) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215; Erratum: Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *119*, 52. (e) Zhao, Y.; Gonzalez-Garcia, N.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 2012; **2006**, *110*, 4942 (E). (f) Zheng, J.; Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 808. (g) Cramer, C. J.; Truhlar, D. G. *Phys. Chem. Chem. Phys.* **2009**, *11*, 10757.

(3) Bento, A. P.; Solà, M.; Bickelhaupt, F. M. *J. Chem. Theory Comput.* **2008**, *4*, 929.

(4) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101/1–18.

(5) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys* **1980**, *58*, 1200.

(6) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785.

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

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

(9) Handy, N. C.; Cohen, A. *Mol. Phys.* **2001**, *99*, 403.

(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. Phys. Chem. A* **2006**, *110*, 13126.

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

(13) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 1849.

(14) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650.

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

(16) Curtiss, L. A.; Redfern, C.; Raghavachari, K.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1999**, *110*, 4703.

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

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

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

(20) van Lenthe, E.; Baerends, E. J. *J. Comput. Chem.* **2003**, *24*, 1142.

(21) te Velde, G.; Bickelhaupt, F. M.; Baerends, E. J.; Fonseca Guerra, C.; van Gisbergen, S. J. A.; Snijders, J. G.; Ziegler, T. *J. Comput. Chem.* **2001**, *22*, 931.

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

(23) Zhao, Y.; Truhlar, D. G. *MN-GSM*, version 4.3; University of Minnesota: Minneapolis, MN, 2009.

(24) Dunlap, B. I. *J. Mol. Struct. (Theochem)* **2000**, *529*, 37.

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

(26) van Gisbergen, S; van Lenthe, E. Scientific Computing & Modelling NV, Amsterdam, the Netherlands. Personal communications, 2010.

(27) Bickelhaupt, F. M. Scheikundig Laboratorium der Vrije Universiteit. Swart, M. Universitat de Girona and Institució Catalana de Recerca i Estudis Avançats (ICREA). Solà, M. Universitat de Girona. Personal communications, 2010.

(28) Liao, M.-S.; Watts, J. D.; Huang, M.-J. *Phys. Chem. Chem. Phys.* **2009**, *11*, 4365.

(29) Swart, M.; Groenhof, A. R.; Ehlers, A. W.; Lammertsma, K. *J. Phys. Chem. A* **2004**, *108*, 5479. de Jong, G. Th.; Bickelhaupt, F. M *J. Chem. Theory Comput.* **2006**, *2*, 322. de Jong, G. Th.; Bickelhaupt, F. M. *J. Phys. Chem. A* **2005**, *109*, 9685. de Jong, G. Th.; Geerke, D. P.; Diefenbach, A.; Solà, M.; Bickelhaupt, F. M. *J. Comput. Chem.* **2005**, *26*, 1006. de Jong, G. Th.; Geerke, D. P.; Diefenbach, A.; Bickelhaupt, F. M. *Chem. Phys.* **2005**, *313*, 261.

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

(31) Ruiz, E.; Salahub, D. R.; Vela, A. *J. Am. Chem. Soc.* **1995**, *117*, 1141.

(32) Hirshfeld, F. L. *Theor. Chem. Acc.* **1977**, *44*, 129.

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

(34) Zhao, Y.; González-García, N.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 2012. 2006, *110*, 4942(E).

CT100082Z

JCTC Journal of Chemical Theory and Computation

# Free Energies of Solvation with Surface, Volume, and Local Electrostatic Effects and Atomic Surface Tensions to Represent the First Solvation Shell

Junjun Liu,[†,‡] Casey P. Kelly,[§] Alan C. Goren,[‡,||] Aleksandr V. Marenich,[§] Christopher J. Cramer,[*,§] Donald G. Truhlar,[*,§] and Chang-Guo Zhan[*,‡]

*Key Laboratory of Pesticide & Chemical Biology of Ministry of Education, College of Chemistry, Central China Normal University, Wuhan 430079, People's Republic of China, Department of Pharmaceutical Sciences, College of Pharmacy, University of Kentucky, 725 Rose Street, Lexington, Kentucky 40536, Department of Chemistry and Supercomputing Institute, 207 Pleasant Street SE, University of Minnesota, Minneapolis, Minnesota 55455, and Division of Natural Sciences & Mathematics, Transylvania University, 300 North Broadway, Lexington, Kentucky 40508*

Received January 13, 2010

**Abstract:** Building on the SVPE (surface and volume polarization for electrostatics) model for electrostatic contributions to the free energy of solvation with explicit consideration of both surface and volume polarization effects, on the SM*x* approach to including first-solvation-shell contributions, and on the linear relationship between the electric field and short-range electrostatic contributions found by Chipman, we have developed a new method for computing absolute aqueous solvation free energies by combining the SVPE method with semiempirical terms that account for effects beyond bulk electrostatics. The new method is called SMVLE, and the elements it contains are denoted by SVPE-CDSL, where SVPE denotes accounting for bulk electrostatic interactions between solute and solvent with both surface and volume contributions, CDS denotes the inclusion of solvent cavitation, changes in dispersion energy, and possible changes in local solvent structure by a semiempirical term utilizing geometry-dependent atomic surface tensions as implemented in SM*x* models, and L represents the local electrostatic effect derived from the outward-directed normal electric field on the cavity surface. The semiempirical CDS and L terms together represent the deviation of short-range contributions to the free energy of solvation from those accounted for by the SVPE term based on the bulk solvent dielectric constant. A solute training set containing a broad range of molecules used previously in the development of SM6 is used here for SMVLE model calibration. The aqueous solvation free energies predicted by the parametrized SMVLE model correlate exceedingly well with experimental values. The square of the correlation coefficient is 0.9949 and the slope is 1.0079. Comparison of the final SMVLE model against the earlier SM*x* solvation model shows that the parametrized SMVLE model not only yields good accuracy for neutrals but also significantly increases the accuracy for ions, making it the best implicit solvation model to date for aqueous solvation free energies of ions. The semiempirical terms associated with the outward-directed electric field account in a physical way for the improvement in the predictive accuracy for ions. The SMVLE method greatly decreases the need to include explicit water molecules for accurate modeling of solvation free energies of ions.

10.1021/ct100025j © 2010 American Chemical Society
Published on Web 03/04/2010

## Introduction

Dielectric continuum solvation models[1,2] have been widely and successfully used for estimating solvation free energies. Such models are also called implicit solvation models because the solvent is not atomistically explicit but rather is implicit in the dielectric medium. In the self-consistent versions of such models, the solvent is considered to be a continuous dielectric medium that is polarized by the solute, leading to a reaction field that in turn polarizes the solute, which changes the solvent polarization, leading ultimately to a self-consistent reaction field (SCRF). A dielectric continuum solvation model accurately describes the long-range permanent-multipole-moment and inductive interactions between solute and bulk solvent; in the language conventionally used in the continuum solvation model literature, both permanent-multipole-moment and induction effects are labeled as electrostatic, and we will follow that convention in the rest of this article. SCRF methods require less computational effort than explicit-solvent approaches involving the same quality treatment of the solute, and this makes them appealing for the study of complex chemical, materials, and biochemical processes and for the rapid screening of many solutes in, for example, molecular docking studies. When the bulk-solvent model is augmented with additional terms to account for the deviation of short-range solute−solvent interactions from the bulk electrostatic model, useful accuracy can be obtained.[3]

One popular way to implement the SCRF approach is to describe the solvent polarization in terms of the electrostatic potential that it introduces inside the solute cavity under the assumption that all solute charge density resides inside the cavity; this is often called the polarized continuum model (PCM).[1,2] However, unconstrained quantum mechanical calculations of solute electronic structure always lead to a tail of the wave function penetrating outside the cavity, thereby causing an additional polarization effect called volume polarization.[4] It has been demonstrated[5] that neglecting charge penetration (also called outlying charge) leads to inconsistencies in the course of solving Poisson's equation. Such inconsistencies render many SCRF implementations sensitive to cavity size[6] and prone to overestimating solvent shifts of energy barriers in aqueous solution.[7,8] In previous studies,[5−7,9,10] a general model called surface and volume polarization for electrostatic interaction (SVPE), or the fully polarizable continuum model (FPCM),[11−18] was developed. This model, implemented for irregularly shaped solute cavities, fully accounts for both surface and volume polarization effects in solute−solvent electrostatic interactions. Therefore, the SVPE solvation model provides a theoretically well justified continuum methodology for studying long-range electrostatic interactions. It has also been useful in a practical sense, having been applied successfully to study mechanisms for various chemical reactions and to make p$K_a$ predictions.[7,8,19,20]

One must bear in mind that absolute solvation free energies result not only from long-range electrostatic interactions between solute and bulk solvent but also from significant short-range contributions, such as short-range and nonbulk electrostatics, as well as cavitation, exchange repulsion, dispersion, and disruption or formation of the nearby solvent structure. (Note that the short-range, nonbulk electrostatic effect may be considered to be a solvent structure effect.) These interactions are not treated satisfactorily within the framework of a pure dielectric continuum model. Previously, the difference between solvation free energies calculated by dielectric continuum solvation methods and experimental solvation free energies has been labeled in an SVPE context as the nondielectric or short-range contribution[21,22] and in other contexts[3] as a cavity-dispersion-solvent-structure (CDS) effect. Such short-range contributions are often[7,8] (but not always[23]) neglected in estimating energy barriers by implicit solvation methods, but for systems with strong hydrophobic effects or hydrogen bonding between solute and solvent molecules, the short-range contributions to the energy barriers may be very significant or even decisive. Furthermore, accounting for the short-range contributions is essential for calculating reliable absolute solvation free energies of neutral solutes.[3]

In a supermolecular approach, the nearby solvent molecules are represented explicitly as components of a cluster continuing the solute. It has been reported that, by employing a combined supermolecule-continuum approach,[15,18,24−27] involving both explicit and implicit solvent, the SVPE method can account for short-range contributions between solute and solvent and hence more accurately predict the absolute solvation free energies for a series of charged chemical species, including $H^+$, $Li^+$, $OH^-$, $e^-$ (the hydrated electron), and $F^-$.[15,18,24,27] The supermolecule-continuum approach explicitly treats a portion of the solvent surrounding the solute at a high quantum mechanical level. The computational efficiency depends heavily on solute size and larger solutes require more solvent water molecules to be explicitly considered, thus in practice limiting the application to smaller solutes. Even more significantly, the proper application of the supermolecule-continuum method requires statistical mechanical averaging over the various possible sites and orientations of the explicit solvent molecules, and this become impractical for even a small number (for example, two or three) of explicit solvent molecules. Finally, one should note that the interaction of solvent molecules in the supersolute (cluster) with the continuum must be treated accurately, for example, through recourse to very high levels of electronic structure theory.

For these reasons, much effort has been made to augment the dielectric continuum model with short-range contributions.[1,2,28] For example, the SM$x$ series of solvation models (with $x$ being 1−6, 8, 8AD, or D)[29−46] augment and correct the bulk electrostatic portion, obtained by the generalized Born (GB) approximation[29,47] (for $x$ = 1−6, 8, or 8AD) or the PCM approximation[2] (for $x$ = 5C or D), with a semiempirical term that accounts for short-range contribu-

* Corresponding authors: e-mail cramer@umn.edu (C.J.C.); truhlar@umn.edu (D.G.T.); zhan@uky.edu (C.-G.Z.).

† Central China Normal University.

‡ University of Kentucky.

§ University of Minnesota.

∥ Transylvania University.

tions. By employing a training data set containing a broad range of solutes, atomic radii used for defining the cavity in electrostatic calculations (such radii are called intrinsic Coulomb radii) were calibrated to calculate the bulk electrostatic interactions, and a set of atomic surface tension parameters was optimized to calculate the short-range contributions. It has been shown[3,43,44] that the accuracy of SM*x* models for predicting absolute aqueous solvation free energies is quite good, about ~0.5 kcal/mol for neutral solutes. Although SM*x* models, like all other solvation models, have larger absolute errors for predicting aqueous solvation free energies of ions, which are much larger than those of neutral solutes, SM*x* still outperforms other continuum models for ionic solvation free energies.[3,48] While the SM*x* models, by implicitly including local electrostatic effects as part of the semiempirical CDS terms, provide significant improvement over PCM models in predicting absolute solvation free energies,[43−45] it is worthwhile to consider more explicit ways to include local electrostatic effects.

The long-range electrostatic contribution, which is a function only of the solvent's bulk dielectric constant, is included in the bulk electrostatic term, but the bulk electrostatic term also includes a somewhat arbitrary approximation to the short-range electrostatic effect, because the solute−solvent boundary that surrounds the solute cavity is located within the region occupied by the first solvation shell, but this shell does not behave like a bulk dielectric. The deviation of short-range electrostatics from bulk electrostatics (this is called the nonbulk electrostatic contribution or the local electrostatic contribution) has previously been included in SM*x* models as a solvent-structure contribution to the CDS term. Here we incorporate a new function, denoted by L for local electrostatics, that treats the nonbulk electrostatic contribution explicitly.

The motivation for the new function is the observation that the short-range electrostatic contribution is linearly correlated with the maximum or minimum outward-directed normal electric field on the solute−solvent boundary surface.[21,22] Therefore we make the contribution beyond the bulk part calculated with SVPE dependent not only on the solvent-accessible atomic surface area of the solute, as in the CDS terms, but also on the outward-directed normal electric field on the cavity surface; the combination of the new local electrostatic terms (L) with the atomic surface tension terms (CDS) is denoted CDSL. Replacing the CDS terms by CDSL terms and replacing the generalized Born approximation (of SM1−SM8 or SM8AD)[3,29−39,41−45] or the PCM approximation (of SM5C and SMD)[40,46] by the SVPE treatment is the essence of the present attempt to make the solvation model more physical. As we have just explained, a long abbreviation for the new method is SVPE-CDSL. It will, however, be more convenient to simply call the new method SMVLE, which denotes solvation model with volume and local electrostatics, since the explicit accounting for volume polarization and local electrostatics are the new elements beyond those included in previous SM*x* solvation models.

The same training set that was used to calibrate parameters for SM6[43] is used here for optimizing the parameters of SMVLE. The absolute aqueous solvation free energies obtained by the parametrized SMVLE model are compared with experimentally measured aqueous solvation free energies to calculate the mean unsigned error (MUE), which measures the predictive accuracy of the solvation model. The accuracy of the SMVLE model is compared with SM6[43] and the recent SM8[44] and SMD[46] solvation models, and the role of the new kind of semiempirical term, called $G_L$, is discussed.

## Methods

**Description of the SMVLE Model.** As explained in the Introduction, the free energy of solvation is a sum of three terms:

$$\Delta G_S^* = \Delta G_{SVPE} + G_{CDS} + G_L \tag{1}$$

Here $\Delta G_S^*$ is the fixed-concentration absolute solvation free energy,[49] $\Delta G_{SVPE}$ is the bulk electrostatic portion calculated by the SVPE method, $G_{CDS}$ is the semiempirical term based on atomic surface tensions, and $G_L$ is the semiempirical electric-field-dependent term, whose form is motivated by Chipman's work on ions[21,22] where two semiempirical terms were generated for anions and cations, respectively. If the standard-state solvation free energy, $\Delta G_S^\circ$, with a concentration corresponding to a solute partial pressure of 1 atm in the gas phase and a solute concentration of 1 M in the liquid phase, is desired instead of $\Delta G_S^*$, then another term, $G_{conc}^\circ =$ 1.89 kcal/mol, must be added to account for the change in concentration.[50] This value of $G_{conc}^\circ$ and all other free energies considered in this paper correspond to a temperature of 298 K. $\Delta G_{SVPE}$ can be expressed as

$$\Delta G_{SVPE} = \langle \Psi^{(1)} | H^0 + \tfrac{1}{2} V | \Psi^{(1)} \rangle - \langle \Psi^{(0)} | H^0 | \Psi^{(0)} \rangle \tag{2}$$

where $\Psi$ is the solute wave function, $H^0$ is the solute Hamiltonian in vacuum, and $V$ is the energy operator associated with the reaction field. The factor of $^1/_2$ in eq 2 stems from assuming a linear response of the surrounding medium to the solute's charge distribution so that half of the induced favorable solute−solvent interaction is canceled by the cost of reorganizing the solvent.[51] The superscripts (0) and (1) refer to the gas-phase isolated molecule and the liquid-phase solution, respectively.

The $G_{CDS}$ term includes free energy changes associated with solvent cavitation, changes in dispersion energy, and possible changes in local solvent structure. It is calculated according to[29,43]

$$G_{CDS} = \sum_{\text{atoms } k} \sigma_k A_k \tag{3}$$

where $A_k$ is the solvent-accessible surface area[52,53] of atom $k$, which depends on the solute geometry, atomic van der Waals radius, and solvent radius, and $\sigma_k$ is the atomic surface tension of atom $k$. The physical basis for eq 3 is that deviations from bulk electrostatics, as well as cavitation, dispersion, and solvent-structural contributions, are all concentrated in the first solvation shell, and $A_k$ is basically a continuous measure of the amount of solvent in the first solvation shell of atom $k$.[29,52,54] The atomic surface tensions are sensitive to local environment, and therefore they are computed according to

$$\sigma_k = \tilde{\sigma}_{Z_k} + \sum_{\text{atoms } k'} \sum_{m=1}^{M} \tilde{\sigma}^{(m)}_{Z_k Z_{k'}} T^{(m)}_{kk'} \tag{4}$$

where $\tilde{\sigma}_{Z_k}$ and $\tilde{\sigma}_{Z_k Z_{k'}}$ are the semiempirical surface tension coefficients for atom $k$ and atom pair $kk'$, and $T_{kk'}$ is a geometry-dependent switching function called a cutoff tanh (referred to as a COT).[34] In most cases the sum over $m$ has only one term, and when $m = 1$, the superscript is omitted. For $k$, $k' = 7$, 6, we have $M = 3$.

The remaining term $G_L$ is motivated by the work of Chipman, who found that more accurate solvation energies could be obtained for ions by adding the following terms to the bulk-electrostatic term:[21,22]

$$G_L \text{ (for anions)} \approx W_{\text{anion}}(E^{\min} - E_0^{\min}) \tag{5}$$

$$G_L \text{ (for cations)} \approx W_{\text{cation}}(E^{\max} - E_0^{\max}) \tag{6}$$

where $G_L$(for anions) and $G_L$(for cations) are "local" (or short-range) electrostatic contributions for anions and cations, respectively; $E^{\min}$ and $E^{\max}$ are the minimum and maximum outward-directed normal electric field on the cavity surface, respectively; and $W_{\text{anion}}$, $E_0^{\min}$, $W_{\text{cation}}$, and $E_0^{\max}$ are fitting parameters. One may interpret these terms as corrections for local electrostatics. Equations 5 and 6 reveal a linear relationship between the local electrostatic contribution and the minimum or the maximum outward-directed normal electric field on the cavity surface for anions and cations, respectively, and the existence of this relationship indicates that a local (short-range) electrostatic interaction between solute and solvent can physically be modeled in term of the outward-directed normal electric field. However, it is not straightforward to generalize the above linear relationships into a formula valid for all solutes, including neutrals and ions with either sign of the charge. For example, one cannot simply combine eqs 5 and 6 because that ignores the local electrostatic effects for neutrals and zwitterions, and therefore the solvation free energy would not vary smoothly along a reaction coordinate where charge is developed or neutralized. The quantities $E^{\min}$ and $E^{\max}$ are also not adequate to represent the local electrostatic effects for dianions or dications with separated charge centers where both the minimum/maximum and the second minimum/maximum normal electric field should be considered. Furthermore, $E^{\min}$ and $E^{\max}$ do not necessarily vary smoothly during geometry optimization. Therefore, a more complicated functional form that does not have these disadvantages is required. We obtain such a function by summing over terms involving the normal electric fields on each surface node and by taking advantage of the properties of the COT function. In particular, we postulate that

$$G_L = \sum_{i=1}^{I} \left\{ B_i \left[ \sum_{m=1}^{M} T(A_i, x_i, -E_m) E_m^{\,i} w_m \right] + B_{i+I} \left[ \sum_{m=1}^{M} T(A_i, x_i, E_m) E_m^{\,i} w_m \right] \right\} \tag{7}$$

$$T(A_i, x_i, E_m) = \frac{1 + \tanh\,[A_i(E_m - x_i)]}{2} \tag{8}$$

where $E_m$ is the outward-directed normal electric field at node $m$ on the cavity surface; $w_m$ is the surface area of node $m$; $M$ is the total number of surface nodes used ($M = 1202$ in the present study); $I$ is an integer that represents the highest power of $E_m$; tanh is the hyperbolic tangent; and $B_i$, $A_i$, and $x_i$ are semiempirical parameters.

In eq 7, the summation over all surface elements means that we consider the local electrostatic contribution not only for ions but also for neutral solutes and zwitterions, and all normal electric fields, instead of only one minimum or one maximum normal electric field for singly charged ions, are considered for all kinds of uncharged and charged systems. In addition, the high powers of $E_m$ allow significant nonlinearity in the relationship. The functional form of eq 7 varies smoothly along a reaction coordinate.

**Computational Details.** The optimized geometries and the corresponding experimental aqueous solvation free energies were obtained from the data set used to calibrate the parameters in the development of SM6,[43] with two exceptions. One exception is that one neutral molecule (*O*-ethyl *O*′-4-bromo-2-chlorophenyl *S*-propyl phosphorothioate or profenofos) in the SM6 training set could not be treated by the SVPE program because of its irregular molecular shape. Thus the SM6 data set has 273 neutral solutes and 143 ions (416 data), and the SMVLE training set has 272 neutral solutes and 143 ions (415 data). All of these molecules and their experimental aqueous solvation free energies are provided as Supporting Information. The second exception concerns the data used for the solvation energies of the ions. Most experimental aqueous solvation free energies are calculated on the basis of thermodynamic cycles, in which the solvation free energies are based on the absolute aqueous solvation free energy of the proton, denoted by $\Delta G_S(H^+)$.[55] The parametrization of SM6 was based on the $\Delta G_S(H^+)$ value of Zhan and Dixon[18] of −264.3 kcal/mol, whereas the later SM8,[44] SM8AD,[45] and SMD[46] were based on the $\Delta G_{\text{sol}}(H^+)$ value of Tissandier et al.[56] of −265.9 kcal/mol. We used the value of Tissandier et al. for the present work.

All the solvation calculations with the SMVLE model were carried out at the HF/6-31+G* electronic structure method.

We previously[43] concluded that the partial charges in some ions are so large that they should be treated by a supermolecule-continuum approach. Therefore, we developed a procedure based on the criterion that if any atom of the ion has partial atomic charge greater than or equal to the partial atomic charge on oxygen in a water molecule, then the ion should be treated as a supermolecule consisting a cluster of a bare ion and one solvent molecule. The ionic data set started with 112 bare ions, and by this criterion, 31 of them should be clustered (so the unclustered instances of these 31 ions are called improperly unclustered). This gives three sets of ions to consider: 81 properly unclustered (PU) ions, 31 improperly unclustered (IU) ions, and 31 monohydrated (MH) ions. The collection of 81 PU and 31 MH ions (total 112 ions) is called the selectively clustered (SC) set; the collection of 81 PU and 31 IU ions (total 112 ions) is called the unclustered (UC) set; and the collection of 81 PU, 31 IU, and 31 MH ions (total 143 ions) is called "all ions". We used all 143 ions set for parametrization, but we give statistics for various subsets for discussion purposes.

The bulk electrostatics were calculated by the SVPE method by using a local version[10] of Gaussian03.[57] The SVPE results depend only on the level and basis set of the quantum mechanical calculation and the isodensity contour value that defines the solute cavity. All the solvation calculations with the SMVLE model were carried out at the HF/6-31+G* level of theory. Previous studies have shown that contour values in the range of 0.0005−0.002 atomic unit lead to a satisfactory description of the electrostatic contributions to the solvation energies for many neutral[6,9] and ionic[21,22] solutes. For this reason, we chose 0.001 atomic unit as the contour value to determine the dielectric cavity. Cavity surface interactions were calculated from a set of 1202 Lebedev grid points and weights that are expected to yield precision of 0.1 kcal/mol or less for SVPE contributions to solvation free energies of all the solutes examined. Surface areas for the CDS term of eq 3 were calculated by the ASA algorithm[58] with the values of Bondi[59] for the atomic radii.

The molecules studied here are generally rigid, except for methyl rotors, whose conformation does not have a large effect on solvation free energies, and consequently no attempt was made to account for relaxation of geometry, change of conformation, or change in vibrational frequencies upon solvation. For solvation free energy calculations, we adopt the Ben-Naim convention[49] that the solute is transferred from a fixed position in the gas phase to a fixed position in solvent (this is called the fixed-concentration solvation free energy above). A value of 78.5 for the dielectric constant value of water is used in all our solvation calculations, which nominally corresponds to 298 K.

**Calibration.** After the bulk electrostatic interactions between solute and solvent were accurately determined for each molecule with the SVPE method, a set of target short-range contributions were obtained from the difference between bulk electrostatic interactions and experimental aqueous solvation free energies. All the CDSL parameters were then subjected to a fitting routine. First the atomic surface tension coefficients were optimized to minimize the root-mean-square error (RMSE) over the 272 neutral solutes. Then the semiempirical parameters in eq 7 were optimized against all 415 molecules (272 neutral solutes and 143 ions) with the atomic surface tension coefficients frozen. The optimization of $A_i$ and $x_i$ was carried out in steps:

**Table 1.** SMVLE Surface Tension Coefficients

| $k$ | $\sigma$ (cal/Å$^2$) | $k$, $k'$ | $\sigma$ (cal/Å$^2$) |
|---|---|---|---|
| H | 57.88 | H, C | −75.22 |
| C | 114.49 | C, C | −70.59 |
| N | −30.82 | H, O | 110.62 |
| O | −84.28 | O, C | 187.69 |
| F | 46.48 | O, O | 98.59 |
| Cl | 14.69 | C, N | 30.94 |
| Br | 12.56 | N, C | −52.83 |
| P | −31.35 | N, C (2)[a] | −261.62 |
| S | −4.13 | N, C (3)[a] | 97.52 |
| | | O, N | 256.52 |
| | | O, P | 79.30 |

[a] Number in parentheses is $m$ when $m$ is not 1.

**Table 2.** Parameters for Local Electrostatic Terms

| | $i = 1$ | $i = 2$ | $i = 3$ |
|---|---|---|---|
| $A_i$ | 1984.0 | 1528.0 | 1488.0 |
| $x_i$ | 0.07 | 0.037 | 0.057 |
| $B_i$ | −2.679 | −23.413 | 453.544 |
| $B_{i+3}$ | 1.454 | −5.64 | −139.35 |

(1) First $I$ was set temporarily to 1, and values from 0.001 to 2 were tried for $x_1$; for each $x_1$, values from 0.1 to 2000 were tried for $A_1$. The values of $x_1$ and $A_1$ that produced the smallest value of the weighted root-mean-square deviation (WRMSD) for the 415 solvation energies in the multiple linear regression fitting process were selected for the next step. In computing the WRMSD in this step and all subsequent steps, the neutral solutes and the ions had relative weights of $W$:1 where $W$ is an integer parameter of the optimization scheme.

(2) $I$ was increased by 1. The values of $A_i$ and $x_i$ already obtained were fixed and the values of $A_I$ and $x_I$ were optimized in the same way as in step 1.

(3) Step 2 was repeated until an arbitrary maximum value of $I$ was reached. In this way, a first-round set of values of $A_i$ and $x_i$ ($i = 1, 2, ..., I$) was obtained.

(4) Now with $I$ fixed at its maximum value, each $A_i$ and $x_i$ was reoptimized with the remaining parameters fixed. For example, $A_1$ and $x_1$ were reoptimized with other $A_i$ and $x_i$ ($i = 2, 3, ..., I$) obtained from previous steps fixed; then the reoptimized values of $A_1$ and $x_1$ along with the values of other $A_i$ and $x_i$ ($i = 3, ..., I$) were fixed in the reoptimization of $A_2$ and $x_2$; and so forth.

(5) Step 4 was repeated until the values of $A_i$ and $x_i$ ($i = 1, 2, ..., I$) did not change.

(6) Steps 1−5 were repeated for several maximal values of $I$. The final value of $I$ was chosen to be 3 as discussed in the next section.

## Results and Discussion

We tested the SMVLE method by examining the errors obtained with various maximal values of $I$ in the range from 1 to 5. We found that the mean errors for ions decreased when $I$ was increased from 1 to 3; however, when $I$ was increased further, the mean errors for the anions improved but those for cations increased. Therefore, we set $I = 3$.

Because solvation free energies of ions are about an order of magnitude larger than those for neutrals, the predictions would have larger errors for ions even if the relative errors were similar. Furthermore, the experimental data for ions usually have larger absolute uncertainties. Another consideration is that our training set contains more neutral data than ionic data. Thus it is a matter of subjective judgment what value to choose for the parametrization weight $W$ to

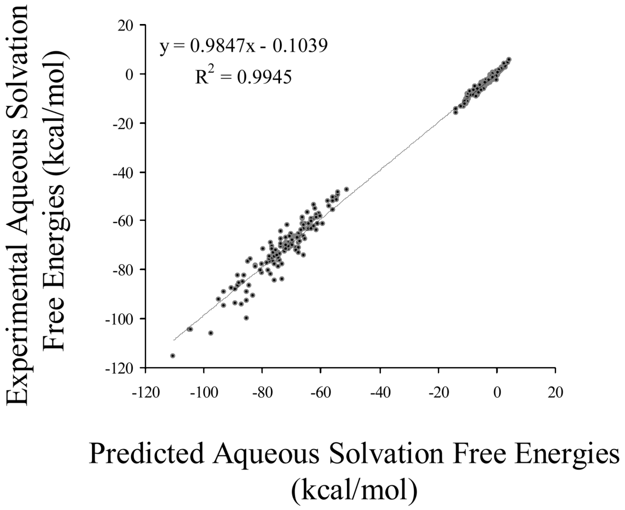


***Figure 1.*** Correlation between the experimental and predicted aqueous solvation free energies.

balance the relative accuracies and desired accuracies for neutrals and for ions; we chose $W = 3$.

The calibrated parameters, namely, the surface tension parameters and the semiempirical $G_L$ parameters, are given in Tables 1 and 2, respectively. The calculated absolute aqueous solvation free energies are obtained by the SMVLE method with these calibrated parameters. They are plotted in Figure 1 along with the experimental values. The predicted aqueous solvation free energies by the parametrized SMVLE are in excellent agreement with the experimental values. The square of the correlation coefficient between calculated and experimental aqueous solvation free energies is 0.9945. Moreover, the slope (0.9847) of the correlation equation is basically 1 and the corresponding constant (0.1039) is nearly 0, showing that the predicted solvation energies by the parametrized SMVLE method are in very good agreement with experimental values.

The mean unsigned errors for each solute class obtained from SMVLE were calculated to compare with those from SM6, SM8, SM8AD, and SMD. For the present study we recalculated the previously reported SM6 and SM8 errors using the reference solvation free energies for water-cluster data (data for ionic clusters and the water dimer) corrected by +2.38 kcal/mol to account for a recently discovered error due to a missing concentration correction term (see ref 46 for more detail) and to convert the SM6 error analysis to the scale based on the proton solvation energy of Tissandier et al.,[56] as discussed above. Note also that although SM8, SM8AD, and SMD were parametrized with 274 data for neutral solutes in water (the 272 data used here plus profenofos and tetramethylsilane), all mean errors given in the present article have been calculated for the 272 neutral data used here. The mean errors for the new SMVLE model and the previous SM6, SM8, SM8AD, and SMD models are shown in Table 3.

The error of the SMVLE model for neutrals is 0.55 kcal/mol and it is close to or better than that of the SM*x* methods, for which mean unsigned errors range between 0.47 (SM6/mPW1PW/6-31G*) and 1.31 (SMD/HF/6-31+G*) kcal/mol (Table 3). The SMVLE method not only retains good accuracy for neutrals but also significantly improves the accuracy for ions. The SMVLE mean unsigned errors for the set of all 143 ions (3.25 kcal/mol) and for the set of 112 UC ions (3.07 kcal/mol) are smaller than the corresponding errors obtained with any other method tested in the present work (and we have shown previously[3,44] that the methods tested here are better than other available methods). The error of the SMVLE model for the set of 112 SC ions is 2.92 kcal/mol, which is close to or better that of the SM*x* methods, for which the MUE ranges between 2.80 (SM8AD/M06-2X/6-31G*) and 4.53 (SMD/mPW1PW/6-31G*) kcal/mol.

The local solvent environment is critical for ions, and it is difficult to simulate with implicit solvent. It was noticed in a previous study[43] that the overall error for aqueous ions decreased when one explicitly bound solvent molecule was introduced. The data for the two subsets of 31 ions (IU and MH) listed in Table 3 show that, in contrast to previous models, explicitly including one solvent water molecule with the ion just slightly increases the predictive accuracy for ions within the SMVLE method. The SMVLE error over 31 tested ions decreases from 4.45 kcal/mol (31 IU ions) to 3.88 kcal/mol (31 MH ions). The difference of ~0.57 kcal/mol in SMVLE is much less than the ~4 kcal/mol difference found on average with the non-SMVLE models tested in the present study (Table 3). The almost identical accuracies obtained with or without the addition of one explicit solvent molecule suggests that SMVLE is capable of successfully modeling the strong local electrostatic interactions between ionic solutes and solvent, and the addition of an explicit solvent molecule is unnecessary for SMVLE.

In an attempt to assess the reason for the success of SMVLE in modeling unclustered ions, we removed the electric field semiempirical term, $G_L$, in eq 1 and we recalculated the errors. The MUE for neutrals changes only about 4%, whereas the MUE for ions increases from ~5 to ~12 kcal/mol. This implies that the improvement in the predictive accuracy for ions can be attributed to the newly introduced $G_L$ term. The values of the outward-directed normal electric fields are acting as indicators of the local solvent environment such that solute−solvent interactions stronger than would be anticipated from the bulk dielectric constant are associated with large values of the outward-directed normal electric field. As a special case, a hydrogen bond between solute and solvent may often be located by the direction of strongest outward-directed normal electric field, and the strength of the hydrogen bond might be represented by the magnitude of the strongest outward-directed normal electric field.[21,22]

SMVLE does not involve the optimization of intrinsic Coulomb radii for each atomic number, as in previous SM*x* models, or for atoms with various bonding types, as in some more empirical models. It is especially encouraging that SMVLE yields good results for ions despite not requiring this. In addition, this is of practical importance because it means SMVLE should be easier to extend to additional atomic numbers, if desired.

Although SMVLE provides significant improvement over all previous SM*x* models for ions, the improvement over SMD is particularly large and especially important. The

***Table 3.*** Mean Unsigned Errors in Aqueous Solvation Free Energies Calculated by SMVLE and Older SM*x* Methods[a]

| | mean unsigned errors (kcal/mol) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ESM | 272 neutrals | 81 PU ions | 31 IU ions | 31 MH ions | 143 all ions | 415 all solutes | 60 SC anions | 52 SC cations | 112 SC ions | 112 UC ions | 384 proper solutes[b] |
| | | | | | SMVLE | | | | | | |
| HF/6-31+G* | 0.55 | 2.55 | 4.45 | 3.88 | 3.25 | 1.48 | 3.17 | 2.63 | 2.92 | 3.07 | 1.24 |
| | | | | | SM6 | | | | | | |
| mPW1PW/6-31G* | 0.47 | 2.55 | 8.57 | 4.34 | 4.24 | 1.77 | 3.45 | 2.58 | 3.05 | 4.22 | 1.22 |
| mPW1PW/6-31+G* | 0.55 | 2.90 | 8.52 | 4.29 | 4.42 | 1.89 | 3.00 | 3.61 | 3.28 | 4.46 | 1.35 |
| | | | | | SM8 | | | | | | |
| mPW1PW/6-31G* | 0.56 | 2.51 | 8.41 | 4.16 | 4.15 | 1.80 | 3.44 | 2.42 | 2.97 | 4.14 | 1.26 |
| mPW1PW/6-31+G* | 0.63 | 2.59 | 8.46 | 4.03 | 4.17 | 1.85 | 2.96 | 3.03 | 2.99 | 4.21 | 1.32 |
| M05-2X/6-31G* | 0.59 | 2.54 | 8.41 | 4.18 | 4.17 | 1.83 | 3.49 | 2.42 | 2.99 | 4.16 | 1.29 |
| | | | | | SM8AD | | | | | | |
| mPW1PW/6-31G* | 0.60 | 2.94 | 6.25 | 2.64 | 3.59 | 1.63 | 2.86 | 2.85 | 2.86 | 3.86 | 1.26 |
| M05-2X/6-31G* | 0.52 | 3.06 | 6.25 | 2.68 | 3.67 | 1.61 | 2.95 | 2.96 | 2.95 | 3.94 | 1.23 |
| M06-2X/6-31G* | 0.61 | 2.88 | 6.05 | 2.59 | 3.50 | 1.61 | 2.78 | 2.82 | 2.80 | 3.76 | 1.25 |
| | | | | | SMD | | | | | | |
| HF/6-31G* | 0.91 | 3.28 | 8.10 | 3.63 | 4.40 | 2.12 | 3.86 | 2.82 | 3.38 | 4.61 | 1.63 |
| HF/6-31+G* | 1.31 | 3.59 | 8.93 | 3.60 | 4.75 | 2.50 | 4.53 | 2.50 | 3.59 | 5.07 | 1.98 |
| mPW1PW/6-31G* | 0.62 | 4.50 | 9.64 | 4.63 | 5.64 | 2.35 | 5.61 | 3.29 | 4.53 | 5.92 | 1.76 |
| M05-2X/6-31G* | 0.59 | 4.08 | 9.11 | 4.24 | 5.21 | 2.18 | 5.01 | 3.09 | 4.12 | 5.47 | 1.62 |
| M06-2X/6-31G* | 0.62 | 4.39 | 9.35 | 4.44 | 5.48 | 2.30 | 5.45 | 3.19 | 4.40 | 5.76 | 1.73 |

[a] ESM = electronic structure method, PU = properly unclustered, IU = improperly unclustered, MH = monohydrated, SC = selectively clustered, UC = unclustered. See text for detailed description of the subsets (Computational Details). [b] All solutes except IU ions.

reason it is especially important is that SMVLE and SMD do not require accurate partial atomic charges, which can sometimes be difficult to obtain for arbitrary levels of theory, extended basis sets, and complex systems. Thus SMVLE and SMD are more generally applicable. We also note that the present test of SMVLE includes diffuse functions (denoted by the "+" in 6-31+G*), and Table 3 shows that in previous methods the use of diffuse functions often decreases accuracy, which we interpreted as due to less stable partial atomic charges in SM6 and SM8 and to outlying charge in SMD. The good performance of SMVLE with a basis set containing diffuse functions is therefore particularly encouraging.

Concerning the computational complexity of the SMVLE method, the ratio of computing time spent for each SCRF cycle of the SMVLE calculation to that for each SCF cycle of the corresponding gas-phase calculation ranges from 1.2 to 1.6 when the number of basis functions used for the solute is larger than ~200.

## Concluding Remarks

We have developed a new method, called SMVLE, for predicting absolute aqueous free energies of solvation by combining (1) the SVPE method, (2) semiempirical atomic surface tensions as used in the SM6 model, and (3) a new functional form, developed in the present study, that explicitly accounts for the local electrostatic effect. The SVPE term accounts for bulk electrostatics; the atomic surface tensions account for solvent cavitation, changes in dispersion energy, and possible changes in local solvent structure; and the final contribution accounts explicitly for nonbulk electrostatics in terms of the local electric field at the solute−solvent boundary. The parameters for SMVLE have been calibrated against a broad range of solutes, including 272 neutrals and 143 ions. The predicted aqueous solvation free energies by the parametrized SMVLE method correlate very well with experiment and have a value of the square of the correlation coefficient equal to 0.9945 and a slope of 0.9847. Comparisons with previous SM*x* solvation models show that the SMVLE model not only has comparable accuracy for neutrals but also impressively increases the predictive accuracy for ions. The semiempirical terms ($G_L$) derived from the electric field are found to be primarily responsible for the increase in predictive accuracy for ions. The outward-directed normal electric fields that make the most important contributions account for strong interactions between the ionic solute and the nearby solvent, which makes the addition of explicit water molecules unnecessary. These encouraging results demonstrate that the parametrized SMVLE is accurate and effective in predicting absolute solvation free energies not only for neutral molecules but also for ions exhibiting strong solute−solvent interactions.

**Acknowledgment.** This work was supported in part by the National Institutes of Health (Grants R01DA013930, R01DA025100, and RC1MH088480 to C.-G.Z.), the National Science Foundation (Grant CHE07-04974 to D.G.T.), and the U.S. Army (Grant US ARMY RES LAB/W911NF09-1-0377 to C.J.C. and D.G.T.). J.L. worked in the laboratory of C.-G.Z. at the University of Kentucky as an exchange graduate student from Central China Normal University. A.G. worked in the laboratory of C.-G.Z. at the University of Kentucky as a visiting professor from Transylvania University. We also acknowledge the Center for Computational Sciences (CCS) at the University of Kentucky for supercomputing time on IBM X-series cluster with 340 nodes or 1360 processors. Computational resources were also provided by Minnesota Supercomputing Institute.

**Supporting Information Available:** Three tables for all molecules involved in this study, including 272 neutral solutes, 112 unclustered ions, and 31 clustered monohydrated ions, and their experimental aqueous solvation free energies. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161–2200.

(2) (a) Tomasi, J.; Persico, M. *Chem. Rev.* **1994**, *94*, 2027–2094. (b) Rivail, J.-L.; Rinaldi, D. In *Computational Chemistry: Reviews of Current Trends*, Vol. 1; Leszczynski, J., Ed.; World Scientific: Singapore, 1996; pp 139−174. (c) Orozco, M.; Luque, F. J. *Chem. Rev.* **2000**, *100*, 4187–4225. (d) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999–3093. (e) *Continuum Solvation Models in Chemical Physics*, Mennucci, B., Cammi, R., Eds.; Wiley: Chichester, U.K., 2007.

(3) (a) Cramer, C. J.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 760–768. (b) Cramer, C. J.; Truhlar, D. G. *Acc. Chem. Res.* **2009**, *42*, 493–497.

(4) (a) Chipman, D. M. *J. Chem. Phys.* **1997**, *106*, 10194–10206. (b) Chipman, D. M. *J. Chem. Phys.* **1996**, *104*, 3276–3289.

(5) Zhan, C.-G.; Bentley, J.; Chipman, D. M. *J. Chem. Phys.* **1998**, *108*, 177–192.

(6) Zhan, C.-G.; Chipman, D. M. *J. Chem. Phys.* **1999**, *110*, 1611–1622.

(7) Zhan, C.-G.; Landry, D. W.; Ornstein, R. L. *J. Phys. Chem. A* **2000**, *104*, 7672–7678.

(8) Chen, X.; Zhan, C.-G. *J. Phys. Chem. A* **2004**, *108*, 6407–6413.

(9) Zhan, C.-G.; Chipman, D. M. *J. Chem. Phys.* **1998**, *109*, 10543–10558.

(10) Vilkas, M. J.; Zhan, C.-G. *J. Chem. Phys.* **2008**, *129*, 194109.

(11) Zhan, C.-G.; Niu, S. Q.; Ornstein, R. L. *J. Chem. Soc., Perkin Trans. 2* **2001**, 23–29.

(12) Dixon, D. A.; Feller, D.; Zhan, C.-G.; Francisco, J. S. *J. Phys. Chem. A* **2002**, *106*, 3191–3196.

(13) Zhan, C.-G.; Dixon, D. A.; Sabri, M. I.; Kim, M. S.; Spencer, P. S. *J. Am. Chem. Soc.* **2002**, *124*, 2744–2752.

(14) Zheng, F.; Zhan, C.-G.; Ornstein, R. L. *J. Phys. Chem. B* **2002**, *106*, 717–722.

(15) Zhan, C.-G.; Dixon, D. A. *J. Phys. Chem. B* **2003**, *107*, 4403–4417.

(16) Zhan, C.-G.; Spencer, P.; Dixon, D. A. *J. Phys. Chem. B* **2003**, *107*, 2853–2861.

(17) Dixon, D. A.; Feller, D.; Zhan, C.-G.; Francisco, J. S. *Int. J. Mass Spectrom.* **2003**, *227*, 421–438.

(18) Zhan, C.-G.; Dixon, D. A. *J. Phys. Chem. A* **2001**, *105*, 11534–11540.

(19) Xiong, Y.; Zhan, C.-G. *J. Org. Chem.* **2004**, *69*, 8451–8458.

(20) (a) Chen, X.; Zhan, C.-G. *J. Phys. Chem. A* **2004**, *108*, 3789–3797. (b) Lu, H.; Chen, X.; Zhan, C.-G. *J. Phys. Chem. B* **2007**, *111*, 10599–10605.

(21) Chipman, D. M. *J. Chem. Phys.* **2003**, *118*, 9937–9942.

(22) Chipman, D. M.; Chen, F. W. *J. Chem. Phys.* **2006**, *124*, 144507.

(23) (a) Cramer, C. J.; Hawkins, G. D.; Truhlar, D. G. *J. Chem. Soc., Faraday Trans.* **1994**, *90*, 1802–1804. (b) Storer, J. W.; Giesen, D. J.; Hawkins, G. D.; Lynch, G. C.; Cramer, C. J.; Truhlar, D. G.; Liotard, D. A. *ACS Symp. Ser.* **1994**, *568*, 24–49. (c) Tuñón, I.; Ruiz-López, M. F.; Rinaldi, D.; Bertrán, J. *J. Comput. Chem.* **1996**, *17*, 148–155. (d) Chuang, Y.-Y.; Cramer, C. J.; Truhlar, D. G. *Int. J. Quantum Chem.* **1998**, *70*, 887–896. (e) Cramer, C. J.; Truhlar, D. G. *Faraday Discuss. Chem. Soc.* **1998**, *110*, 477–479. (f) Chuang, Y.-Y.; Radhakrishnan, M. L.; Fast, P. L.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **1999**, *103*, 4893–4909. (g) Chuang, Y.-Y.; Truhlar, D. G. *J. Am. Chem. Soc.* **1999**, *121*, 10157–10167. (h) Jaque, P.; Marenich, A.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. C* **2007**, *111*, 5783–5799. (i) Kim, Y.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **2009**, *113*, 9109–9114. (j) Kim, Y.; Marenich, A. V.; Zheng, J.; Kim, K. H.; Kołodziejska-Huben, M.; Rostkowski, M.; Paneth, P.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 59–67.

(24) Zhan, C.-G.; Dixon, D. A. *J. Phys. Chem. A* **2004**, *108*, 2020–2029.

(25) Xiong, Y.; Zhan, C.-G. *J. Phys. Chem. A* **2006**, *110*, 12644–12652.

(26) Zhan, C.-G.; Landry, D. W.; Ornstein, R. L. *J. Am. Chem. Soc.* **2000**, *122*, 2621–2627.

(27) Zhan, C.-G.; Dixon, D. A. *J. Phys. Chem. A* **2002**, *106*, 9737–9744.

(28) (a) Rivail, J. L.; Terryn, B.; Rinaldi, D.; Ruiz-Lopez, M. F. *J. Mol. Struct. THEOCHEM* **1985**, *120*, 387–400. (b) Rinaldi, D.; Costa Cabral, B. J.; Rivail, J.-L. *Chem. Phys. Lett.* **1986**, *125*, 495–499.

(29) Cramer, C. J.; Truhlar, D. G. *J. Am. Chem. Soc.* **1991**, *113*, 8305–8311.

(30) Cramer, C. J.; Truhlar, D. G. *J. Comput. Chem.* **1992**, *13*, 1089–1097.

(31) Cramer, C. J.; Truhlar, D. G. *Science* **1992**, *256*, 213–217.

(32) Chambers, C. C.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 16385–16398.

(33) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem.* **1996**, *100*, 19824–19839.

(34) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **1998**, *102*, 3257–3271.

(35) Li, J. B.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *Chem. Phys. Lett.* **1998**, *288*, 293–298.

(36) Zhu, T. H.; Li, J. B.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Phys.* **1998**, *109*, 9117–9133.

(37) Zhu, T. H.; Li, J. B.; Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Phys.* **1999**, *111*, 5624–5624.

(38) Li, J. B.; Zhu, T. H.; Hawkins, G. D.; Winget, P.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. *Theor. Chem. Acc.* **1999**, *103*, 9–63.

(39) Li, J. B.; Zhu, T. H.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **2000**, *104*, 2178–2182.

(40) Dolney, D. M.; Hawkins, G. D.; Winget, P.; Liotard, D. A.; Cramer, C. J.; Truhlar, D. G. *J. Comput. Chem.* **2000**, *21*, 340–366.

(41) Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 6532–6542.

(42) Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. *Theor. Chem. Acc.* **2005**, *113*, 107–131 .

(43) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2005**, *1*, 1133–1152.

(44) Marenich, A. V.; Olson, R. M.; Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 2011–2033.

(45) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 2447–2464.

(46) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2009**, *113*, 6378–6396.

(47) (a) Hoijtink, G. J.; Deboer, E.; Vandermeij, P. H.; Weijland, W. P. *Recl. Trav. Chim. Pays-Bas Belg.* **1956**, *75*, 487–503. (b) Peradejordi, F. *Cah. Phys.* **1963**, *17*, 393. (c) Constanciel, R.; Contreras, R. *Theor. Chim. Acta* **1984**, *65*, 1–11. (d) Tucker, S. C.; Truhlar, D. G. *Chem. Phys. Lett.* **1989**, *157*, 164–170. (e) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129. (f) Cramer, C. J.; Truhlar, F. G. *Rev. Comput. Chem.* **1995**, *6*, 1–72. (g) Bashford, D.; Case, D. A. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129–153.

(48) Cramer, C. J.; Truhlar, D. G. *Acc. Chem. Res.* **2009**, *42,* 493.

(49) Ben-Naim, A. *J. Phys. Chem.* **1978**, *82*, 792–803.

(50) Cramer, C. J.; Truhlar, D. G. In *Free Energy Calculations in Rational Drug Design*; Reddy, M. R., Erion, M. D., Eds.; Kluwer: New York, 2001; pp 63−95.

(51) Cramer, C. J.; Truhlar, D. G. In *Solvent Effects and Chemical Reactivity*; Tapia, O., Bertran, J., Eds.; Kluwer: Dordrecht, The Netherlands, 1996; pp 1−81.

(52) Hermann, R. B. *J. Phys. Chem.* **1972**, *76*, 2754–2759.

(53) Lee, B.; Richards, F. M. *J. Mol. Biol.* **1971**, *55*, 379−400.

(54) Nemethy, G.; Scheraga, H. A. *J. Chem. Phys.* **1962**, *36*, 3401–3417.

(55) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2006**, *110*, 16066–81.

(56) Tissandier, M. D.; Cowen, K. A.; Feng, W. Y.; Gundlach, E.; Cohen, M. H.; Earhart, A. D.; Coe, J. V.; Tuttle, T. R. *J. Phys. Chem. A* **1998**, *102*, 7787–7794.

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

(58) Liotard, D. A.; Hawkins, G. D.; Lynch, G. C.; Cramer, C. J.; Truhlar, D. G. *J. Comput. Chem.* **1995**, *16*, 422–440.

(59) Bondi, A. *J. Phys. Chem.* **1954**, *58*, 929–939.

CT100025J

JCTC Journal of Chemical Theory and Computation

# A Test to Evaluate the Performance of Aromaticity Descriptors in All-Metal and Semimetal Clusters. An Appraisal of Electronic and Magnetic Indicators of Aromaticity

Ferran Feixas,[†] J. Oscar C. Jiménez-Halla,[†] Eduard Matito,[‡] Jordi Poater,[†] and Miquel Solà*,[†]

*Institut de Química Computacional and Departament de Química, Universitat de Girona, Campus de Montilivi, 17071 Girona, Catalonia, Spain and Institute of Physics, University of Szczecin, 70-451 Szczecin, Poland*

Received January 19, 2010

**Abstract:** As compared to classical organic aromatic compounds, the evaluation of aromaticity in all-metal and semimetal clusters is much more complex. For a series of these clusters, it is frequently found that different methods used to discuss aromaticity lead to divergent conclusions. For this reason, there is a need to evaluate the reliability of the different descriptors of aromaticity to provide correct trends in all-metal and semimetal aromatic clusters. This work represents the first attempt to assess the performance of aromaticity descriptors in all-metal clusters. To this end, we introduce the series of all-metal and semimetal clusters $[X_nY_{4-n}]^{q\pm}$ (X, Y = Al, Ga, Si, and Ge; $n = 0-4$) and $[X_nY_{5-n}]^{4-n}$ (X = P and Y = S and Se; $n = 0-5$) with predictable aromaticity trends. Aromaticity, in these series, is quantified by means of nucleus-independent chemical shifts (NICS) and electronic multicenter indices (MCI). Results show that the expected trends are generally better reproduced by MCI than by NICS. It is found that NICS(0)$_\pi$ is the kind of NICS that performs better among the different NICS indices analyzed.

## 1. Introduction

The discovery of aromaticity in $Al_4^{2-}$,[1] an all-metal inorganic cluster, in 2001 by Boldyrev, Wang, and co-workers fuelled the interest for the study of all-metal and semimetal inorganic clusters with aromatic properties (for three recent reviews see ref 2). At variance with the classical aromatic organic molecules that possess only $\pi$-electron delocalization, the aromaticity in inorganic clusters is more complex due to the peculiarities of chemical bonding in metal systems. These metal compounds present $\sigma$-, $\pi$-, and $\delta$-[3] or even $\phi$-[4]electron delocalization, thus, giving rise to the so-called multifold aromaticity/antiaromaticity[2,5] as well as cases of conflicting aromaticity.[2a,6]

Most of the methods to quantify aromaticity have been developed for the classical aromatic organic molecules, and they cannot be applied to inorganic clusters without further reinvestigation. This is the case, among others, of the harmonic oscillator model of aromaticity (HOMA)[7] or the aromatic fluctuation (FLU)[8] indicators of aromaticity that take benzene, the paradigmatic aromatic molecule, or other aromatic organic molecules as a reference in their definitions. Likewise, resonance energies (RE) or aromatic stabilization energies (ASE)[9] are very difficult to compute accurately in all-metal clusters because of the lack of appropriate reference systems.[5c,10] For the moment, the most widely used methods to discuss aromaticity in all-metal clusters have been the simple electron counting based on the $4n + 2$ Hückel's rule[11] and the calculation of the nucleus-independent chemical shifts (NICS).[12] Less common is the use of electronic multicenter indices (MCI),[13] for which few examples can be found in the literature.[14]

[*] Corresponding author: Tel.: +34.972.41.89.12; Fax: +34.972.41.83.56; E-mail address: miquel.sola@udg.edu.

[‡] University of Szczecin.

[†] Universitat de Girona.

10.1021/ct100034p © 2010 American Chemical Society
Published on Web 03/18/2010

Although the $4n + 2$ rule affords the simple test of aromaticity, electron counting alone does not provide always direct evidence of aromaticity/antiaromaticity.[2c,15] For instance, $Al_4^{2-}$ contains one pair of delocalized $\pi$-electrons and two pairs of $\sigma$-electrons that contribute to the overall aromaticity of this species.[1,6a,16] The two $\pi$-electrons obey the $4n + 2$ Hückel's rule for monocyclic's $\pi$-systems. Although this is not the case for the four $\sigma$-electrons, it was found that the two pairs of delocalized $\sigma$ electrons belong to molecular orbitals (MOs) that follow orthogonal radial and tangential directions, which makes them totally independent,[17] thus, separately following the $4n + 2$ rule. This is a clear example that simple total electronic counts sometimes lead to erroneous conclusions.[2c,15] Similarly, in planar polycyclic boron clusters, it has been found that the aromaticity is not related to the total number of $\pi$-electrons.[18]

Probably the most widely employed method to analyze the aromaticity of all-metal species is the NICS index. This descriptor, proposed by Schleyer and co-workers[12] as a magnetic index of aromaticity, is a valuable indicator of aromaticity that is used by many researchers. It is defined as the negative value of the absolute shielding computed at a ring center or at some other interesting point of the system, usually 1 Å above the ring center. Rings with large negative NICS values are considered aromatic. The more negative the NICS values, the more aromatic the rings are. Nonaromatic species have NICS values close to zero, and positive NICS values are indicative of antiaromaticity. Recently, dissected NICS techniques based on the analysis of individual canonical MOs contributions to NICS[19] have been successfully applied to analyze multifold and conflicting aromaticity/antiaromaticity in all-metal clusters.[20]

In a recent work, some of us reported that NICS profiles calculated in the perpendicular direction of each ring are useful to classify all-metal and semimetal clusters into three groups according to their aromatic, nonaromatic, or antiaromatic character.[21] In addition, Tsipis has recently demonstrated that the NICS$_{zz}$-scan patterns, along with symmetry-based selection rules, can unequivocally probe the antiaromaticity in a wide range of antiaromatic organic and inorganic rings/cages.[22] We also showed[21] that single-point NICS calculations fail to provide correct trends for some particular systems. For example, we found unexpectedly that $C_{2v}$ $GeAl_3^-$ is more aromatic than $D_{4h}$ $Al_4^{2-}$, according to NICS(0) values.[21] In another work, we discovered that the NICS and MCI predicted changes of aromaticity in $Mg_3^{2-}$ when coordinated to alkalimetal cations follow opposite trends.[14c] Similar results were reported by Chattaraj et al. for the metal complexation of $Al_4^{2-}$.[14b] The reason for the divergence between NICS and MCI is unclear in some cases. The connection between these indicators is not obvious because NICS values are computed as a response to an external magnetic field, and virtual orbitals are involved in the calculation, while in the computation of electron sharing indices (ESI), such as MCI, only occupied orbitals are used. In fact, it is well-known that a delocalized system is a necessary but not a sufficient condition to have a ring current.[14a,23]

In a subsequent work,[24] we introduced a series of 15 aromaticity tests that can be used to analyze the advantages and the drawbacks of a group of aromaticity descriptors. Based on the results obtained for a set of 10 indicators of aromaticity, including NICS and MCI, we concluded that MCI were the most accurate among all indices examined in that work.[24] In addition, the fact that the $\pi$-component of the four center-electron index in $Al_4^{2-}$ is almost the same as that of $C_4H_4^{2+}$ seems to indicate an apparent good behavior of MCI for all-metal clusters.[14a] The 15 tests of aromaticity proposed in the previous work[24] involved only classical aromatic molecules, having expected aromaticity trends based on accumulated chemical experience. In the present work, we introduce a new test containing several inorganic all-metal clusters with two main aims: first, to investigate the performance of NICS and MCI to provide expected aromaticity trends in all-metal clusters; and second, to analyze whether, among different NICS definitions, there is a particular NICS index that performs consistently better than the rest.

To this end, we have chosen the four-membered ring (4-MR) series of valence isoelectronic inorganic species $[X_nY_{4-n}]^{q\pm}$ (X, Y = Al, Ga, Si, and Ge; $n = 0-4$) that have a predictable trend of aromaticity. Thus, one can predict a steep decrease in aromaticity when going from $Al_4^{2-}$ to, for instance, $GeAl_3^-$ due to the reduction of symmetry and to the substitution of one Al atom by a more electronegative Ge atom. A smooth reduction of aromaticity when going from $Al_3Ge^-$ to $Al_2Ge_2$ is also likely, although more arguable. And the same should occur from $Ge_4^{2+}$ to $Al_2Ge_2$. Therefore, for instance, the expected order of aromaticity in one (X = Al and Y = Ge) of the six series chosen is $Al_4^{2-} > Al_3Ge^- \geq Al_2Ge_2 \leq AlGe_3^+ < Ge_4^{2+}$. A similar behavior is likely to be present in a series where X and Y come from different groups of the Periodic Table. In the series with X = Al and Ga and Y = Si and Ge, the electronegativity of X and Y is significantly different and, consequently, large changes in bond distances and angles are observed. On the other hand, when X and Y belong to the same group (for instance, X = Al and Y = Ga), electronegativity and geometrical parameters remain almost unchanged. This fact leads to small changes in the aromaticity and, thus, the expected trend becomes $Al_4^{2-} \geq Al_3Ga^{2-} \sim Al_2Ga_2^{2-} \sim AlGa_3^{2-} \leq Ga_4^{2-}$. Finally, apart from these series, we have included two $[X_nY_{5-n}]^{4-n}$ (X = P and Y = S and Se; $n = 0-5$) series of 5-MRs. It is worth noting that the electronic, molecular structure, and aromaticity of some of the systems studied here have been analyzed in previous theoretical and experimental works ($Al_4^{2-}$,[1,5c,6a,10,14a,b,15a,16,17,23,25] $Ga_4^{2-}$,[10,25a,26] $Al_3Si^-$,[27] $Al_2Si_2$,[5f,25a,28] $AlSi_3^+$,[28b] $Si_4^{2+}$,[5d,28b] $Al_3Ge^-$,[27a] $Al_2Ge_2$,[5f,28a] $AlGe_3^+$,[28b] $Ge_4^{2+}$,[28b] $Ga_3Si^-$,[29] $Ga_2Si_2$,[5f,25a,28] $GaSi_3^+$,[28b] $Ga_3Ge^-$,[29] $Ga_2Ge_2$,[5f,28a] $GaGe_3^+$,[28b] $Si_2Ge_2^{2+}$,[28b] and $P_5^-$ [5e,30]). It has been found that the elements of the $[X_nY_{4-n}]^{q\pm}$ series present $\sigma$- and $\pi$-aromaticity, while $[X_nY_{5-n}]^{4-n}$ compounds are only $\pi$-aromatic. The present study complements the available experimental and theoretical data, which is scarce or missing for some clusters, and enables a systematic analysis of aromaticity trends along the eight clusters series $[X_nY_{4-n}]^{q\pm}$ (X, Y = Al, Ga, Si, and Ge;

$n = 0-4$) and $[X_nY_{5-n}]^{4-n}$ (X = P and Y = S and Se; $n = 0-5$), all obtained with the same methodology.

## 2. Computational Details

All calculations in this work were performed by means of the Gaussian03[31] computational package. The gas-phase optimized geometries reported here were calculated in the framework of density functional theory (DFT) using the B3LYP functional,[32] which combines the three-parameter Becke's exchange[33] and Lee−Yang−Parr's correlation[34] nonlocal functionals. The 6-311+G(d) basis set[35] was used for all calculations.

In the present work, we report results for two unstable dianions such as $Al_4^{2-}$ and $Ga_4^{2-}$. In a recent work, Lambrecht et al.[36] have shown that $Al_4^{2-}$ is unstable as compared to $Al_4^-$ + free $e^-$ and, consequently, its properties change significantly when increasing the number of diffuse functions in the basis set. Indeed, after inclusion of certain number of diffuse functions, the $Al_4^{2-}$ evolves to $Al_4^-$ + free $e^-$. In this sense, Lambrecht and co-workers[36] warned about the validity of calculations carried out for such unstable dianions. In a recent comment[37] (see also the rebuttal in ref 38) on the work by Lambrecht et al.,[36] Zubarev and Boldyrev argued against this point of view and considered that the bound state of the individual $Al_4^{2-}$ is an adequate model of $Al_4^{2-}$ in a stabilizing environment, such as in $NaAl_4^-$ or $Na_2Al_4$. They also considered that calculations for isolated $Al_4^{2-}$ species using a 6-311+G(d) basis provide an accurate model for the $Al_4^{2-}$ unit embedded in a stabilizing environment. Following the Zubarev and Boldyrev arguments,[37] we will discuss the properties of the bound state in these two metastable dianions by employing the 6-311+G(d) basis set.

NICS values were computed also with the B3LYP/6-311+G(d) method through the gauge-including atomic orbital method (GIAO)[39] implemented in Gaussian03. The magnetic shielding tensor was calculated for the ghost atoms located at the ring centers (NICS(0)) determined by the nonweighted mean of the heavy atom coordinates and also for the ring critical point (RCP), the point of lowest density in the ring plane,[40] as suggested by Cossío et al.,[41] to yield NICS(0)$^{rcp}$ values. In some high-symmetry molecules both points, the ring center and the RCP, coincide. In addition, NICS has been also calculated at 1 Å above the molecular plane of the ring (NICS(1) and NICS(1)$^{rcp}$).[42] NICS(1) is considered to better reflect the $\pi$-electron effects than that of NICS(0). The out-of-plane tensor component of NICS (NICS(0)$_{zz}$, NICS(0)$_{zz}^{rcp}$, NICS(1)$_{zz}^{rcp}$, and NICS(1)$_{zz}$) have also been collected. This latter quantity gives probably the best measure of aromaticity among the different NICS-related definitions in organic molecules.[24,43] Moreover, NICS(0)$_\pi$, NICS(0)$_\pi^{rcp}$, NICS(0)$_\sigma$, and NICS(0)$_\sigma^{rcp}$ have been obtained from the decomposition of NICS into its MO components using the NBO 5.0 program.[44] For these calculations, only the contributions of valence orbitals have been taken into account. The dissected NICS methods have already been widely applied to analyze multifold aromaticity in all-metal clusters.[20]

For the aromaticity analysis, we have also applied the MCI.[13b,c] MCI is a particular extension of the $I_{ring}$ index:[13a]

$$I_{ring}(\mathbf{A}) = \sum_{i_1,i_2,...,i_N} n_{i_1}, ..., n_{i_N} S_{i_1i_2}(A_1)S_{i_2i_3}(A_2), ..., S_{i_Ni_1}(A_N) \tag{1}$$

$n_i$ is the occupancy of MO $i$ and $S_{ij}(A)$ is the overlap between MOs $i$ and $j$ within the molecular space assigned to atom $A$. Summing up all the $I_{ring}$ values resulting from the permutations of indices $A_1$, $A_2$, ..., $A_N$, the mentioned MCI index[13c] is defined as

$$\mathrm{MCI}(\mathbf{A}) = \frac{1}{2N}\sum_{P(\mathbf{A})} I_{ring}(\mathbf{A}) \tag{2}$$

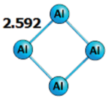
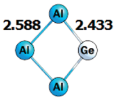
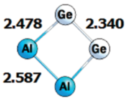
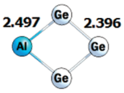
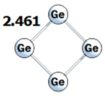
where $P(\mathbf{A})$ stands for a permutation operator which interchanges the atomic labels $A_1$, $A_2$, ..., $A_N$ to generate up to the $N$! permutations of the elements in the string $\mathbf{A}$.[13b,45] MCI and $I_{ring}$ give an idea of the electron sharing between all atoms in the ring. The more positive the MCI values,[13c,46] the more aromatic the rings. For planar species, as those treated in the present work, $S_{\sigma\pi}(A) = 0$ and both MCI and $I_{ring}$ can be exactly split into the $\sigma$- and $\pi$-contributions. This feature is especially interesting to evaluate multifold aromaticity in all-metal clusters. Finally, although several atomic partitions may be used for the calculations of the overlap between MOs $i$ and $j$ within the molecular space assigned to atom $A$,[14a,47] we have chosen in the present work the partition carried out in the framework of the quantum theory of atoms-in-molecules (QTAIM) of Bader,[40,48] by which atoms are defined from the condition of zero-flux gradient in the one-electron density, $\rho(\mathbf{r})$. Calculation of atomic overlap matrices (AOM) and computation of MCI have been performed with the AIMPAC[49] and ESI-3D[50] collection of programs.[51] Since MCI and $I_{ring}$ yield very similar results, in this work, we report only MCI values.

## 3. Results and Discussion

In this section, we first discuss the series $[Al_nGe_{4-n}]^{2-n}$ ($n = 0-4$) in detail. Then the results for the rest of the $[X_nY_{4-n}]^{q\pm}$ series are briefly analyzed. Finally, the molecular structure and aromaticity of the $[X_nY_{5-n}]^{4-n}$ (X = P and Y = S and Se; $n = 0-5$) clusters are presented.

Table 1 contains the molecular structure and the MCI and NICS results obtained for the series $[Al_nGe_{4-n}]^{2-n}$ ($n = 0-4$). The number of valence electrons is 14 for all members of the series. The molecular structure of the ground state of $Al_4^{2-}$ and $Ge_4^{2+}$ clusters is $D_{4h}$ square planar.[28b] The geometry of the ground state of $Al_3Ge^-$ [27a] and $AlGe_3^+$ [28b] clusters is a planar distorted rhombus of $C_{2v}$ symmetry. It is worth noting that in the case of heteroatoms having substantially different electronegativities, as in $Al_3C^-$, the most stable molecular structure becomes the $C_{3v}$ pyramidal.[27a] This can be likely attributed to the loss of aromaticity due to higher electronegativity differences. For $Al_2Ge_2$, there are two possible planar structures corresponding to the *cis* and *trans* configurations.[5f,28a] According to previous studies, the $Al_2Ge_2$ cluster is more stable in the *trans* configuration,[5f,28b] while the *cis* is the most stable configuration for $Ga_2Ge_2$,[5f,28b] $Al_2Si_2$,[5f,25a,28b] and $Ga_2Si_2$.[5f,28b] In order to make comparisons between different series easier,

***Table 1.*** Molecular Structure and Values of the MCI, $MCI_\pi$, and $MCI_\sigma$ (in electrons) and the NICS (in ppm) Indices[a]

| | $Al_4^{2-}$ | $Al_3Ge^-$ | $Al_2Ge_2$ | $AlGe_3^+$ | $Ge_4^{2+}$ |
|---|---|---|---|---|---|
| |  |  |  |  |  |
| Symmetry | $D_{4h}$ | $C_{2v}$ | $C_{2v}$ | $C_{2v}$ | $D_{4h}$ |
| MCI | 0.356 | 0.206 | 0.165 | 0.171 | 0.386 |
| $MCI_\pi$ | 0.187 | 0.114 | 0.102 | 0.128 | 0.187 |
| $MCI_\sigma$ | 0.169 | 0.092 | 0.063 | 0.043 | 0.199 |
| NICS(0) | -34.45 | -34.57 | -35.00 | -38.18 | -42.20 |
| $NICS(0)^{rcp}$ | -34.45 | -31.98 | -30.00 | -31.90 | -42.20 |
| NICS(1) | -27.39 | -26.97 | -26.56 | -28.72 | -30.28 |
| $NICS(1)^{rcp}$ | -27.39 | -25.66 | -24.16 | -25.73 | -30.28 |
| $NICS(0)_{zz}$ | -66.14 | -64.09 | -64.68 | -67.84 | -69.36 |
| $NICS(0)_{zz}^{rcp}$ | -66.14 | -62.75 | -62.88 | -66.10 | -69.36 |
| $NICS(1)_{zz}$ | -54.85 | -53.66 | -53.65 | -57.31 | -58.56 |
| $NICS(1)_{zz}^{rcp}$ | -54.85 | -52.59 | -52.11 | -55.36 | -58.56 |
| $NICS(0)_\pi$ | -21.72 | -20.95 | -20.32 | -22.71 | -24.31 |
| $NICS(0)_\pi^{rcp}$ | -21.72 | -19.82 | -18.02 | -19.40 | -24.31 |
| $NICS(0)_\sigma$ | -12.23 | -12.62 | -13.36 | -13.79 | -16.23 |
| $NICS(0)_\sigma^{rcp}$ | -12.23 | -11.25 | -10.78 | -10.94 | -16.23 |

[a] Calculated at the ring center and at the ring critical point (RCP) for the series $Al_4^{2-}$, $Al_3Ge^-$, $Al_2Ge_2$, $AlGe_3^+$, and $Ge_4^{2+}$ at the B3LYP/6-311+G(d) level of theory.

in all cases, we have taken the *cis* configuration for the $X_2Y_2^{q\pm}$ species, despite, in some cases, the *trans* configuration is the most stable. Aromatic ring current shielding (ARCS) results from Jusélius et al. indicate that the aromaticity of the *cis* and *trans* configurations is similar in $Al_2Si_2$.[25a]

The MCI and $MCI_\pi$ values obtained for $Al_4^{2-}$ are 0.356 and 0.187 e, respectively, not far from the values, 0.341 and 0.161 e, reported by Mandado et al.[14a] with the same QTAIM partition of the molecular space and with the same B3LYP/6-311+G(d) methodology.[52] The value of the $MCI_\pi$ can be easily and analytically obtained for any ring $X_4$ of $D_{4h}$ symmetry with only two $\pi$-electrons occupying the same orbital, such as in $Al_4^{2-}$. Following the procedure that we applied to get analytical delocalization indices for two $\pi$-electron cyclic systems,[53] in $Al_4^{2-}$ there is a single $\pi$-orbital involved in the sum of eq 1, and the self-overlap of this $\pi$-orbital in a given basin is by construction $S_{\pi\pi}(A) = 1/4$. Application of eq 2 yields a MCI value of $3/16 = 0.1875$ e, which is independent of the basis set used for the calculation (it can differ only if correlated wave functions are used).[47a] On the other hand, Mandado et al.[14a] and Roy et al.[14b] reported total MCI values of 0.335 and 0.313 e, respectively, both calculated with the same methodology used in the present work but using Mulliken instead of QTAIM partition. This shows that the effect of using different partitions for the calculation of MCI is minor in the case of $Al_4^{2-}$. Our results also point out that the $\pi$ delocalization in the $Al_4^{2-}$ species is slightly larger than that of the $\sigma$ (0.187 vs 0.169 e). This is in line with previous dissected NICS results,[14a] showing that $NICS(0)_\pi$ is somewhat more negative than $NICS(0)_\sigma$ and also showing the result from the electronic localization function indicating higher $\pi$- than $\sigma$-aromaticity in $Al_4^{2-}$,[25b] but in contrast with the fact that the ring current in $Al_4^{2-}$ has a negligible contribution from the two $\pi$-electron system.[16,23] For symmetry reasons, the $MCI_\pi$ values of $D_{4h}$ $Al_4^{2-}$ and $Ge_4^{2+}$ clusters with two $\pi$-electrons are exactly the same, 0.187 e. Total MCI and absolute NICS values are somewhat larger for $Ge_4^{2+}$, but this is, in part, the result of shorter Ge−Ge bond lengths. For comparison purposes, let us add here that the MCI and NICS(0) values for $D_{4h}$ $C_4H_4^{2+}$,

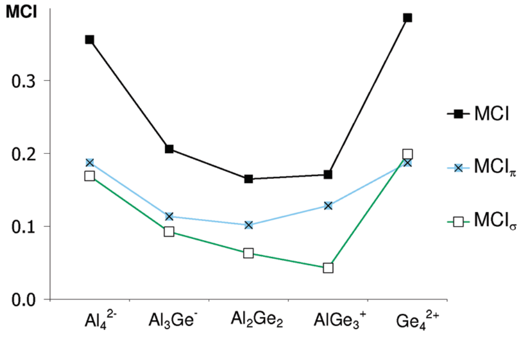


***Figure 1.*** Variation of MCI, $MCI_\pi$, and $MCI_\sigma$ (in electrons) along the series $Al_4^{2-}$, $Al_3Ge^-$, $Al_2Ge_2$, $AlGe_3^+$, and $Ge_4^{2+}$.

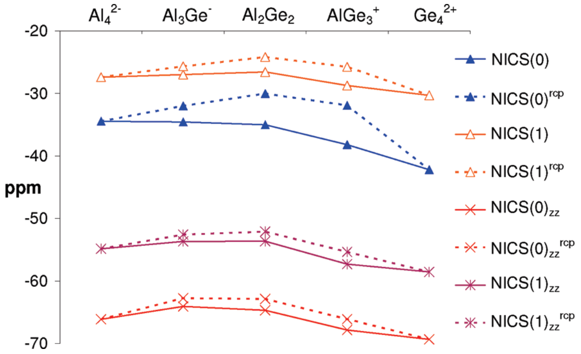


***Figure 2.*** Comparison between NICS (ppm) indices calculated at the ring center and at the RCP (dotted line) along the series $Al_4^{2-}$, $Al_3Ge^-$, $Al_2Ge_2$, $AlGe_3^+$, and $Ge_4^{2+}$.

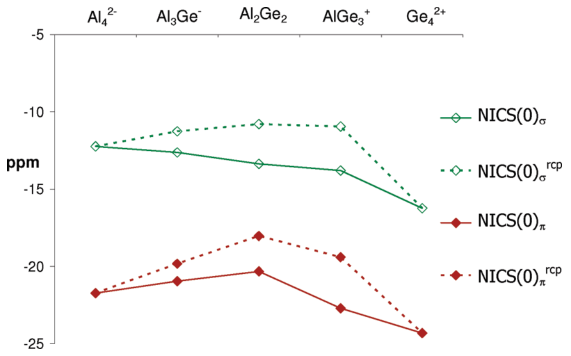


***Figure 3.*** Comparison between dissected NICS (ppm) indices calculated at the ring center and at the RCP (dotted line) along the series $Al_4^{2-}$, $Al_3Ge^-$, $Al_2Ge_2$, $AlGe_3^+$, and $Ge_4^{2+}$.

the organic molecule being most comparable to $Al_4^{2-}$ or $Ge_4^{2+}$, are 0.185 e and −15.62 ppm, respectively.

Figures 1 and 2 depict the trends observed along the series $Al_4^{2-}$ to $Ge_4^{2+}$ for the MCI and NICS indices, respectively. Interestingly, both total MCI and $MCI_\pi$ curves have a clear concave ∪ shape providing the expected order of aromaticity, i.e., $Al_4^{2-} > Al_3Ge^- \geq Al_2Ge_2 \leq AlGe_3^+ < Ge_4^{2+}$. As to the NICS values, NICS(0) fails showing a steady increase of aromaticity along the $Al_4^{2-}$ to $Ge_4^{2+}$ series. Remarkably, $NICS(0)^{rcp}$ yields the anticipated order of aromaticity, indicating that the point selected to compute the NICS value in inorganic clusters may have a relevant influence in the aromaticity trends obtained. The distance between the geometrical ring center and the RCP is 0.19, 0.28, and 0.39 Å in $Al_3Ge^-$, $Al_2Ge_2$, and $AlGe_3^+$, respectively, showing significant differences of −2.58, −4.99, and −6.28 ppm between the NICS(0) and $NICS(0)^{rcp}$. Clearly, the RCP yields better results than the geometrical center of the ring for NICS(0). Minor changes in the out-of-plane component of the NICS(0) and $NICS(0)_{zz}$, computed in the ring center or in the RCP, were found. In both cases, $NICS(0)_{zz}$ and $NICS(0)_{zz}^{rcp}$, the shape of the curve is close to the expected one with the only exception of the aromaticity of $Al_3Ge^-$, that is found to be slightly lower than that of $Al_2Ge_2$. Moreover, NICS(1), $NICS(1)^{rcp}$, $NICS(1)_{zz}$, and $NICS(1)_{zz}^{rcp}$ curves show the expected ∪ behavior. These results oppose to previous claims asserting that NICS(0) is better suited than NICS(1) for the evaluation of aromaticity in all-metal clusters.[5f]

Finally, dissected NICS calculations have been performed in order to study the trends of π- and σ-aromaticity (see Figures 3 and 4). Interestingly, $NICS(0)_\pi$ and $NICS(0)_\sigma$ calculated at the ring center show opposite trends. The first reproduces the predicted concave shape, while the latter exhibits a smooth increase of aromaticity from $Al_4^{2-}$ to

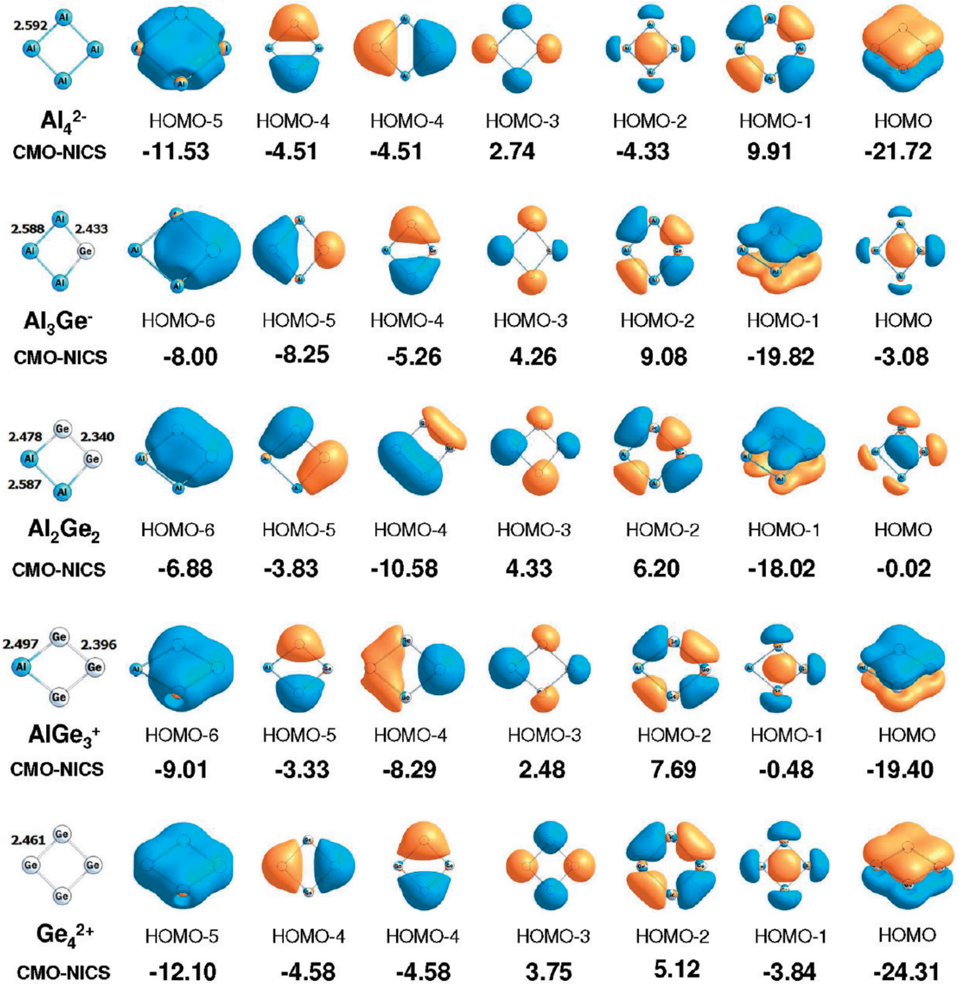


***Figure 4.*** Canonical molecular orbital contribution to $NICS(0)^{rcp}$ (ppm) for the series $Al_4^{2-}$, $Al_3Ge^-$, $Al_2Ge_2$, $AlGe_3^+$, and $Ge_4^{2+}$.

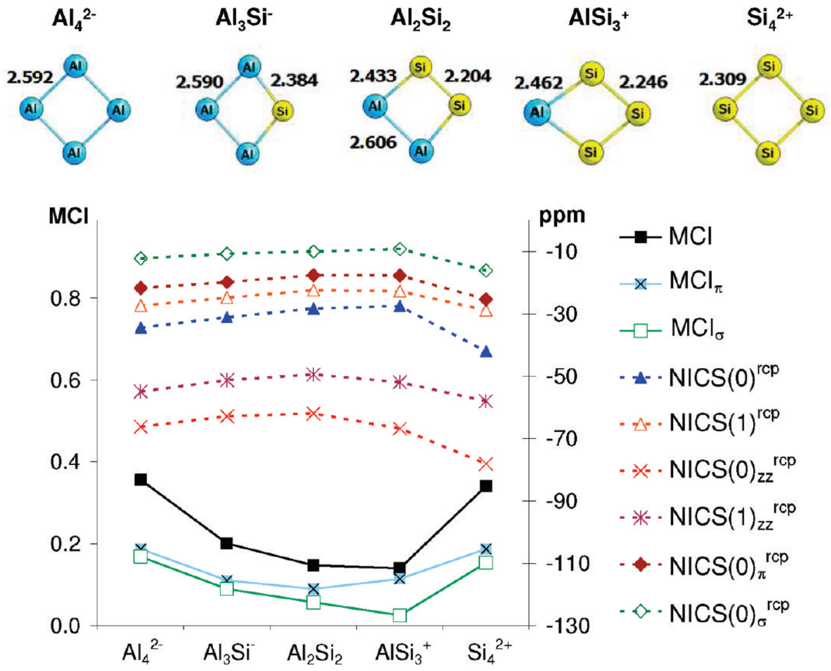


***Figure 5.*** Variation of MCI, $MCI_\pi$, and $MCI_\sigma$ (in electrons) and NICS (ppm) indices calculated at the RCP (dotted line) along the series $Al_4^{2-}$, $Al_3Si^-$, $Al_2Si_2$, $AlSi_3^+$, and $Si_4^{2+}$.

$Ge_4^{2+}$. This last behavior has been previously observed for NICS(0). Therefore, in this series, the $\sigma$ contribution is responsible for the NICS(0) failing. On the other hand, as previously seen, $NICS(0)_\sigma^{rcp}$ provides the expected order, indicating that $NICS(0)_\sigma$ is more sensitive than $NICS(0)_\pi$ at the point where the NICS is computed. Figures 2 and 3 show the differences between the NICS measured at the ring center and at the RCP. In all cases, the latter reproduces properly the expected ∪ shape. Consequently, here after, only the NICS indices calculated at the RCP will be taken into account. Moreover, in the next series, only the figures with selected MCI and NICS curves will be presented (see Supporting Information for tables with the complete set of results). Figure 4 depicts the valence orbitals with its individual contribution to NICS. Interestingly, the radial $\sigma$-orbital (HOMO-2 in $Al_4^{2-}$) reproduces the predicted concave shape, as does the total $NICS_\sigma^{rcp}$, while the tangential $\sigma$-orbital (HOMO-1 in $Al_4^{2-}$) presents a different behavior. In the rest of this work, the individual contributions will only be used to explain the cases where $NICS_\sigma$ or $NICS_\pi$ fail.

Next, the remaining series where X and Y belong to different groups of the Periodic Table are briefly analyzed. The molecular structure, the MCI and NICS curves obtained for the 4-MR valence isoelectronic series $[Al_nSi_{4-n}]^{2-n}$ ($n = 0-4$) are depicted in Figure 5. Whereas MCI indicates somewhat larger aromaticity for $Al_4^{2-}$, all NICS indices give $Si_4^{2+}$ as the most aromatic cluster. The trend obtained when going from $Al_4^{2-}$ to $Si_4^{2+}$ for all the studied indicators of total ($\sigma + \pi$) aromaticity is the same and corresponds to the expected one, except for $NICS(0)^{rcp}$ and MCI that yield $AlSi_3^+$ slightly less aromatic than $Al_2Si_2$. When the $\sigma-\pi$ separation is applied to the MCI and $NICS(0)^{rcp}$ indices, it is found that $MCI_\pi$ and $NICS(0)_\pi^{rcp}$ show the expected ∪ shape, while $MCI_\sigma$ and $NICS(0)_\sigma^{rcp}$ constantly decrease when going from $Al_4^{2-}$ to $AlSi_3^+$. Then the $\sigma$-aromaticity abruptly increases from $AlSi_3^+$ to the full symmetric $Si_4^{2+}$. As it will be seen in the next series, $MCI_\sigma$ and the absolute value of $NICS(0)_\sigma^{rcp}$ tend to decrease when group 13 atoms (Al, Ga) are substituted by group 14 atoms (Si, Ge), except when the $D_{4h}$ structure is reached. In contrast to the previous series, the contribution of the radial $\sigma$-orbital to NICS increases when going from $X_2Y_2$ to $XY_3^+$ (see Figure 6). This fact leads to a slightly lower $\sigma$ aromaticity in $XY_3^+$ than $X_2Y_2$. However, this effect is, in most cases, canceled out by the $\pi$ contribution when total MCI and NICS indices are analyzed, and consequently, the expected concave shape is observed.

After that we consider the two valence isoelectronic series $[X_nY_{4-n}]^{2-n}$ (X = Ga and Y = Si and Ge; $n = 0-4$). Figures 7 and 8 depict the most relevant MCI and NICS curves obtained. Interestingly, almost all indices coincide in giving a similar aromaticity to $Ga_4^{2-}$ and $Ge_4^{2+}$ clusters, while MCI differs from NICS in giving somewhat larger aromaticity to $Ga_4^{2-}$ than to $Si_4^{2+}$. In addition, all indices provide the expected ∪ shape, except $NICS(0)^{rcp}$, $NICS(0)_\sigma^{rcp}$, and $MCI_\sigma$ for the series $[Ga_nSi_{4-n}]^{2-n}$ ($n = 0-4$) that, as before, yield $GaSi_3^+$, slightly less aromatic than $Ga_2Si_2$. In the case of MCI, the decrease of $\sigma$-aromaticity in $GaSi_3^+$ is counteracted by $MCI_\pi$. In general, small differences between $X_2Y_2$ and $XY_3^+$ species are observed. For such cases, one cannot rule out the possibility that a change of method or basis set may lead to a different order of aromaticity.

Figure 9 collects the molecular structure, the MCI and the NICS values obtained for the valence isoelectronic series $[Al_nGa_{4-n}]^{2-}$ ($n = 0-4$). In comparison with the previous series, changes in bond distances and angles are now smaller due to successive substitution of Al by Ga. The B3LYP/6-311+G(d) Ga−Ga bond distance of 2.568 Å found is not far from the 2.618 Å obtained at the CCSD(T)/6-311+G(d) level of theory[26b] and from the 2.47 Å found in an organometallic compound synthesized by

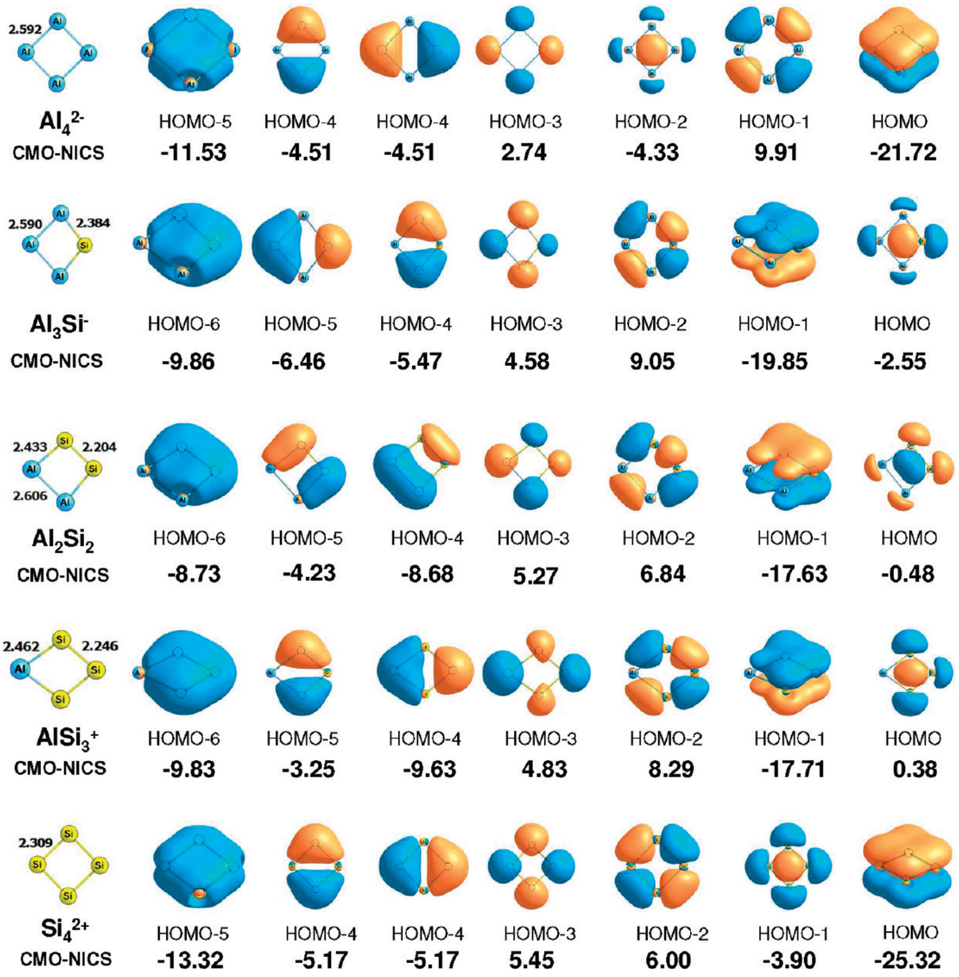


***Figure 6.*** Canonical molecular orbital contribution to NICS(0)$^{rcp}$ (ppm) for the series $Al_4^{2-}$, $Al_3Si^-$, $Al_2Si_2$, $AlSi_3^+$, and $Si_4^{2+}$.

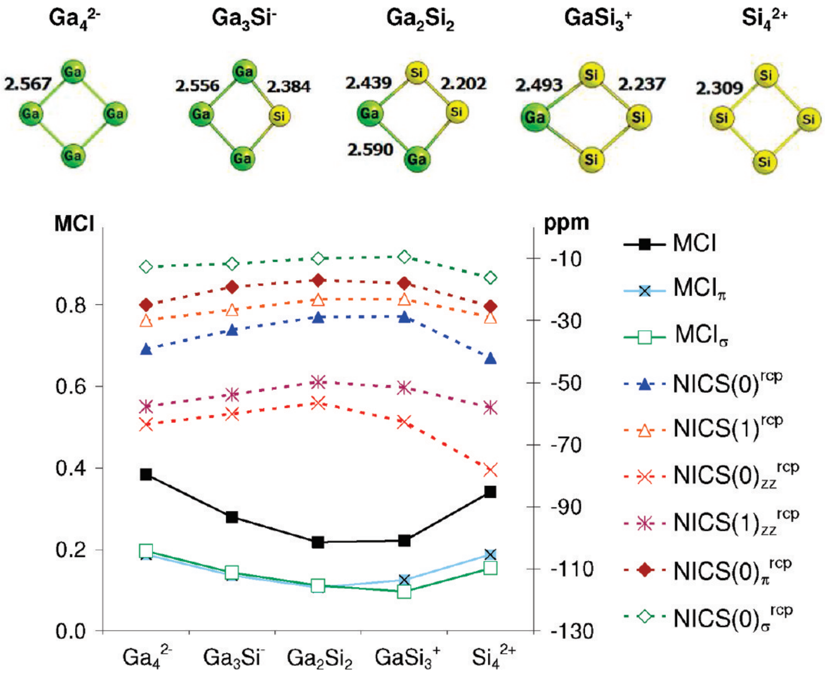


***Figure 7.*** Variation of MCI, $MCI_\pi$, and $MCI_\sigma$ (in electrons) and NICS (ppm) indices calculated at the RCP (dotted line) along the series $Ga_4^{2-}$, $Ga_3Si^-$, $Ga_2Si_2$, $GaSi_3^+$, and $Ga_4^{2+}$.

Twamley and Power[26a] that contains the $Ga_4^{2-}$ unit coordinated to two $K^+$ and bound diagonally to two phenyl carbons. Likewise, changes in aromaticity along the $Al_4^{2-}$ to $Ga_4^{2-}$ series according to MCI and NICS results are relatively small. This is not unexpected taking into account that Ga and Al belong to group 13 and that the differences

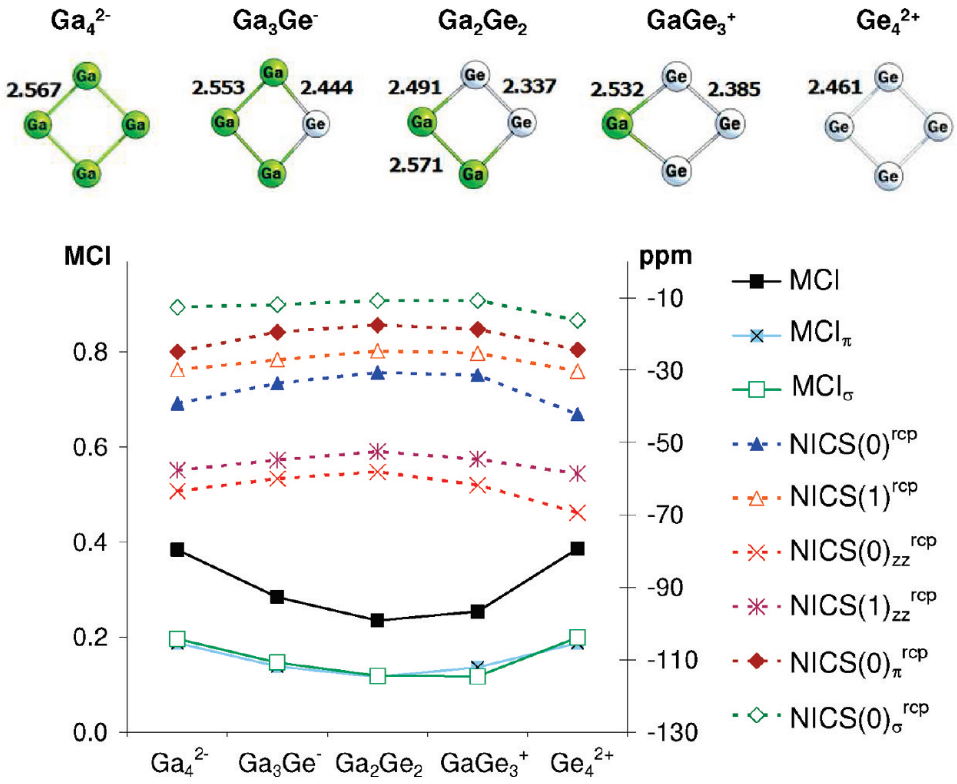


***Figure 8.*** Variation of MCI, MCI$_\pi$, and MCI$_\sigma$ (in electrons), and NICS (ppm) indices calculated at the RCP (dotted line) along the series $Ga_4^{2-}$, $Ga_3Ge^-$, $Ga_2Ge_2$, $GaGe_3^+$, and $Ge_4^{2+}$.

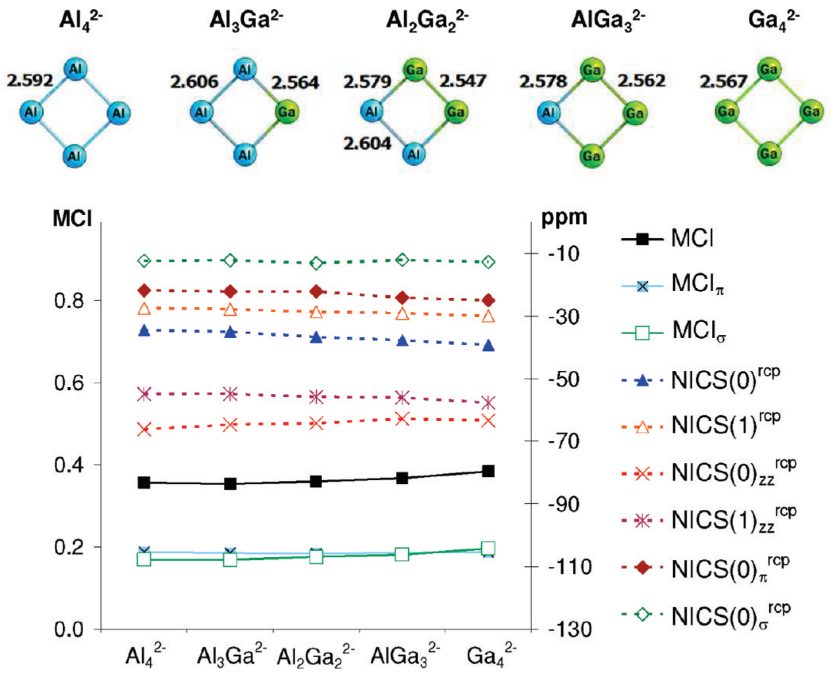


***Figure 9.*** Variation of MCI, MCI$_\pi$, and MCI$_\sigma$ (in electrons) and NICS (ppm) indices calculated at the RCP (dotted line) along the series $Al_4^{2-}$, $Al_3Ga^{2-}$, $Al_2Ga_2^{2-}$, $AlGa_3^{2-}$, and $Ga_4^{2-}$.

of electronegativity between them are lower than those between Al and Ge or Si. MCI yields $Al_4^{2-}$ slightly less aromatic than $Ga_4^{2-}$ in disagreement with a crude evaluation of the resonance energy of $Na_2Al_4$ and $Na_2Ga_4$ by Boldyrev and Kuznetsov.[10] These authors reported that the resonance energy of $Na_2Al_4$ compound is higher than that of the $Na_2Ga_4$ cluster by about 10 kcal·mol$^{-1}$. Correspondingly, NICS results predict higher aromaticity for $Ga_4^{2-}$, except in the case of NICS(0)$_{zz}^{rcp}$. The trends of NICS(0)$^{rcp}$, NICS(1)$^{rcp}$, NICS(1)$_{zz}^{rcp}$, and NICS(0)$_\pi^{rcp}$ in Figure 9 show a steady increase of aromaticity from $Al_4^{2-}$ to $Ga_4^{2-}$. MCI$_\pi$ curve has a clear ∪ shape, although it is less pronounced than in the previous series, providing the expected order of aromaticity, i.e., $Al_4^{2-} \geq Al_3Ga^{2-} \sim Al_2Ga_2^{2-} \sim AlGa_3^{2-} \leq Ga_4^{2-}$. Finally, MCI and NICS(0)$_{zz}^{rcp}$ yield the correct trend except for $Al_2Ga_2^{2-}$

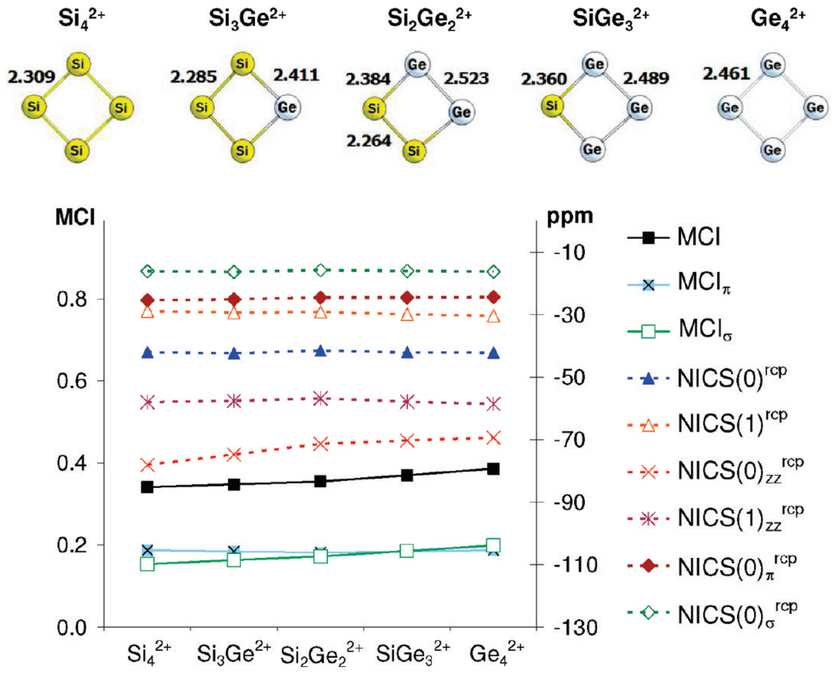


***Figure 10.*** Variation of MCI, MCI$_{\pi}$, and MCI$_{\sigma}$ (in electrons) and NICS (ppm) indices calculated at the RCP (dotted line) along the series $Si_4^{2+}$, $Si_3Ge^{2+}$, $Si_2Ge_2^{2+}$, $SiGe_3^{2+}$, and $Ge_4^{2+}$.

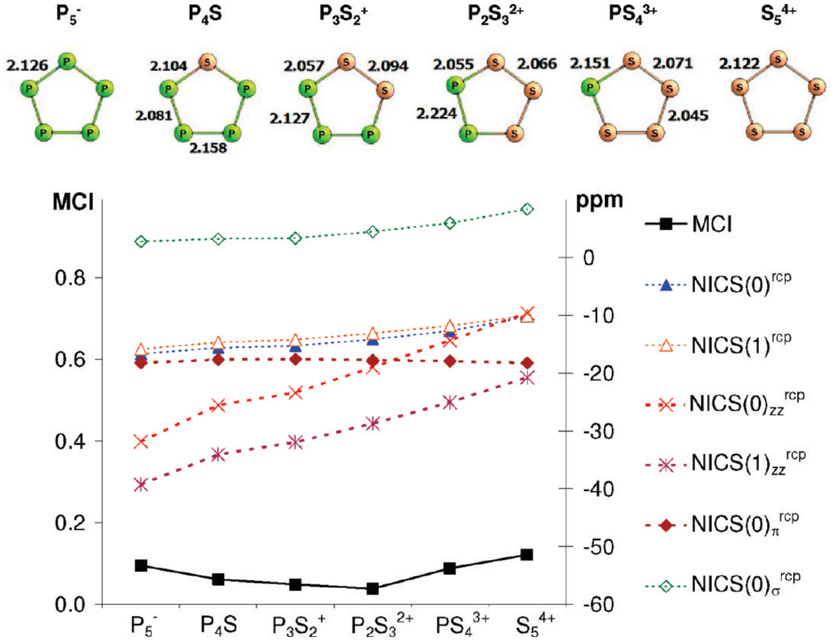


***Figure 11.*** Variation of MCI, MCI$_{\pi}$, and MCI$_{\sigma}$ (in electrons) and NICS (ppm) indices calculated at the RCP (dotted line) along the series $P_5^-$, $P_4S$, $P_3S_2^+$, $P_2S_3^{2+}$, $PS_4^{3+}$, and $S_5^{4+}$.

that is found to be slightly more aromatic than $Al_3Ga^{2-}$ (MCI) or $AlGa_3^{2-}$ (NICS(0)$_{zz}^{rcp}$).

For the valence isoelectronic series $[Si_nGe_{4-n}]^{2+}$ ($n$ = 0−4), we have a similar situation as in the series $Al_4^{2-}$ to $Ga_4^{2-}$. Thus, changes in bond distances and angles are small by successive substitution of Si by Ge atoms (see Figure 10). Likewise, changes in aromaticity along the $Si_4^{2+}$ to $Ge_4^{2+}$ series according to MCI and NICS results are generally minor. This is attributed again to the fact that Si and Ge belong to the same group 14 and that the electronegativities of Si and Ge are almost the same. All methods predict a slightly higher aromaticity for $Ge_4^{2+}$ as compared to $Si_4^{2+}$, except in the case of NICS(0)$_{zz}^{rcp}$ and NICS(0)$_{\pi}^{rcp}$. As to the trends shown in Figure 10, only for MCI$_{\pi}$ and NICS(1)$_{zz}^{rcp}$, the curves have a clear concave shape providing the expected order of aromaticity, i.e., $Si_4^{2+} \geq Si_3Ge^{2+} \sim Si_2Ge_2^{2+} \sim SiGe_3^{2+} \leq Ge_4^{2+}$. MCI indicates a steady increase of aromaticity along the $Si_4^{2+}$ to $Ge_4^{2+}$ series, while NICS(0)$_{zz}^{rcp}$ gives exactly the opposite trend. Finally, NICS(0)$^{rcp}$, NICS(1)$^{rcp}$, and NICS(0)$_{\sigma}^{rcp}$ show a tendency to increase from $Si_4^{2+}$ to $Ge_4^{2+}$ but present some oscillatory behavior. Results show that it is more difficult to observe a clear trend of aromaticity in the series with elements that belong to the

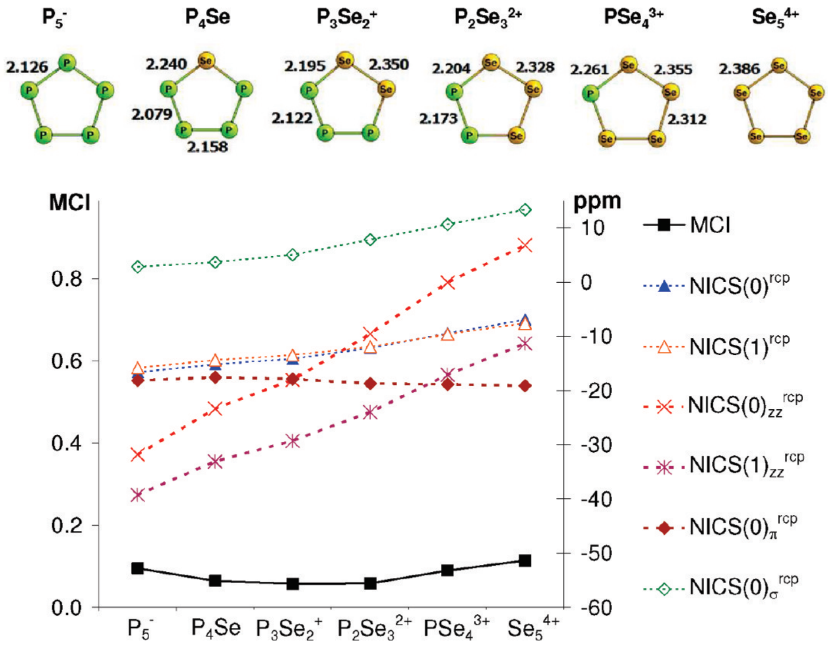


***Figure 12.*** Variation of MCI, $MCI_\pi$, and $MCI_\sigma$ (in electrons) and NICS (ppm) indices calculated at the RCP (dotted line) along the series $P_5^-$, $P_4Se$, $P_3Se_2^+$, $P_2Se_3^{2+}$, $PSe_4^{3+}$, and $Se_5^{4+}$.

***Table 2.*** Summary of the Results Obtained at the B3LYP/6-311+G(d) Level for the Six Series Studied with Seven Descriptors of Aromaticity Analyzed

| series | MCI | $MCI_\pi$ | $NICS(0)^{rcp}$ | $NICS(1)^{rcp}$ | $NICS(0)_{zz}^{rcp}$ | $NICS(1)_{zz}^{rcp}$ | $NICS(0)_\pi^{rcp}$ |
|---|---|---|---|---|---|---|---|
| Al/Ge | yes | yes | yes | yes | unclear[a] | yes | yes |
| Al/Si | unclear[a] | yes | unclear[a] | yes | yes | yes | yes |
| Ga/Si | yes | yes | unclear[a] | unclear[a] | yes | yes | yes |
| Ga/Ge | yes | yes | yes | yes | yes | yes | yes |
| P/S | yes | yes | no | no | no | no | yes |
| P/Se | yes | yes | no | no | no | no | unclear[a] |

[a] Fails only in ordering one molecule (see text).

same group of the Periodic Table. In these series, aromaticity remains basically unchanged by successive substitution. Due to the lack of a well-defined trend, these last two series should not be used as possible tests to evaluate the performance of aromaticity indices in all-metal clusters. Still, results indicate presumably, that the $MCI_\pi$ index performs better than the rest for these two series.

Finally, we discuss the $[X_nY_{5-n}]^{4-n}$ (X = P and Y = S and Se; $n = 0-5$) series of 5-MR clusters. Figures 11 and 12 list the molecular structure, MCI, and NICS curves for these valence isoelectronic clusters. Since the $D_{5h}$ rings of $P_5^-$, $S_5^{4+}$, and $Se_5^{4+}$ have six π-electrons, it is not possible to derive the $MCI_\pi$ value using simple algebra without computation. Among these series, $P_5^-$ is the only inorganic cluster that has been studied previously with quantum mechanical methods.[5e,30] These works show that $D_{5h}$ $P_5^-$ possess six π-electrons in three π molecular orbitals, resulting in π-aromaticity according to the $4n + 2$ Hückel's rule. In line with this view, the MCI and $MCI_\pi$ values of $P_5^-$ differ by only 0.001 e, indicating that the contribution of the σ-electrons to the total MCI value is irrelevant. The same is true for all the members of these two series. MCI values yield more aromatic $S_5^{4+}$ and $Se_5^{4+}$ clusters than $P_5^-$, while all NICS indices predict the opposite behavior. From the trends depicted in Figures 11 and 12, it is found that both MCI and $MCI_\pi$ curves present the expected ∪ shape ($MCI_\pi$ is not represented in Figures 11 and 12 because it coincides almost exactly with MCI). However, all NICS values yield a continuous reduction of aromaticity along the series $P_5^-$ to $S_5^{4+}$ and to $Se_5^{4+}$. Remarkably, $NICS(0)_\pi^{rcp}$ differs from the rest of NICS indices, showing the same behavior of MCI and $MCI_\pi$. $NICS(0)_\pi^{rcp}$ values yield more aromatic $S_5^{4+}$ and $Se_5^{4+}$ clusters than $P_5^-$ and provide a clear ∪ shape, with the only exception of $P_4Se$ being a little less aromatic than $P_3Se_2^+$. On the other hand, σ-orbitals are responsible for the reduction of aromaticity that is shown by the rest of NICS indices along the series. For comparison purposes, the MCI and NICS(0) values for $D_{5h}$ $C_5H_5^-$, the most similar organic molecule to $P_5^-$, are 0.0704 e and −15.63 ppm, respectively.

## 4. Concluding Remarks

The summary of the results obtained for six of the series analyzed can be found in Table 2. In this table, we write "yes" when a certain index follows the expected trend in aromaticity for a given series, "no" otherwise, and "unclear" when the failure of the index is minor (for instance, the index falls short only for the ordering of one species in a given series). The series analyzed constitute a new test to evaluate the performance of descriptors of aromaticity in the exciting field of all-metal clusters.

Results in Table 2 indicate that the multicenter indices perform generally better than NICS, especially the $\mathrm{MCI}_\pi$. Our results reinforce the superior behavior of $\mathrm{NICS}(0)_\pi{}^{\mathrm{rcp}}$ as compared to $\mathrm{NICS}(0)^{\mathrm{rcp}}$, $\mathrm{NICS}(1)^{\mathrm{rcp}}$ and their corresponding out-of-plane components. Indeed, $\mathrm{NICS}(0)_\pi{}^{\mathrm{rcp}}$ gives the correct trends for all studied species, except for the relative aromaticity of $P_4Se$ in comparison with $P_3Se_2{}^+$, and for the valence isoelectronic series $[Al_nGa_{4-n}]^{2-}$ and $[Si_nGe_{4-n}]^{2+}$ ($n = 0-4$). These two latter series, for which only the $\mathrm{MCI}_\pi$ index yields the correct trend, have not been included in Table 2 because aromaticity results show that there is not a well-defined trend along these series. The fact that $\mathrm{NICS}(0)_\pi{}^{\mathrm{rcp}}$ performs better than $\mathrm{NICS}(0)^{\mathrm{rcp}}$, $\mathrm{NICS}(1)^{\mathrm{rcp}}$, or their corresponding out-of-plane components for the $[X_nY_{4-n}]^{q\pm}$ (X, Y = Al, Ga, Si, and Ge; $n = 0-4$) clusters is somewhat disturbing given the fact that these molecules display both $\sigma$- and $\pi$-aromaticity. The reason must be found in the $\mathrm{NICS}(0)_\sigma{}^{\mathrm{rcp}}$ component that fails to account for the expected order of aromaticity. Remarkably, NICS values in inorganic aromatic clusters strongly depend on the point where they are calculated. Thus, while $\mathrm{NICS}(0)^{\mathrm{rcp}}$ provides the expected trend, NICS(0) fails predicting a steady increase when going from $Al_4{}^{2-}$ to $Ge_4{}^{2+}$.

NICS and MCI are indices of aromaticity that do not require reference values and, consequently, they are likely the most useful indicators of aromaticity for all-metal and semimetal clusters. The present study indicates that if one wants to order a series of inorganic compounds according to their aromaticity, it is recommendable to use multicenter electronic indices or $\mathrm{NICS}(0)_\pi{}^{\mathrm{rcp}}$ values. However, for this purpose neither NICS(0) nor NICS(1) are reliable. On the other hand, if one only wants to discuss whether a given cluster is aromatic or not, then both MCI and NICS, and particularly NICS-scan, do a good job to classify all-metal and semimetal clusters into aromatic, nonaromatic, and antiaromatic.[21]

Finally, the performance of NICS and MCI has been validated for the light atoms of Periodic Table, but still remain to be assessed for more complicated transition metals having $\delta$- or $\phi$-electron delocalization. More research is underway in our laboratory concerning this particular issue.

**Acknowledgment.** Financial help has been furnished by the Spanish MICINN project no. CTQ2008-03077/BQU and by the Catalan DIUE through project no. 2009SGR637. M.S. is indebted to the Catalan DIUE for financial support through the ICREA Academia Prize 2010. F.F. and J.P. thank the MICINN for the doctoral fellowship no. AP2005-2997 and the Ramón y Cajal contract, respectively. E.M. acknowledges financial support from the Lundbeck Foundation Center and from Marie Curie IntraEuropean Fellowship, Seventh Framework Programme (FP7/2007−2013), under grant agreement no. PIEF-GA-2008-221734. We thank the Centre de Supercomputació de Catalunya (CESCA) for partial funding of computer time.

**Supporting Information Available:** B3LYP/6-311+G(d) optimized Cartesian coordinates, the complete set of tables and dissected NICS for all the inorganic systems discussed in this work are available. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Li, X.; Kuznetsov, A. E.; Zhang, H.-F.; Boldyrev, A.; Wang, L.-S. *Science* **2001**, *291*, 859–861.

(2) (a) Boldyrev, A. I.; Wang, L.-S. *Chem. Rev.* **2005**, *105*, 3716–3757. (b) Tsipis, C. A. *Coord. Chem. Rev.* **2005**, *249*, 2740–2762. (c) Zubarev, D. Y.; Averkiev, B. B.; Zhai, H.-J.; Wang, L.-S.; Boldyrev, A. I. *Phys. Chem. Chem. Phys.* **2008**, *10*, 257–267.

(3) (a) Zhai, H. J.; Averkiev, B. B.; Zubarev, D. Y.; Wang, L.-S.; Boldyrev, A. I. *Angew. Chem., Int. Ed.* **2007**, *46*, 4277. (b) Averkiev, B. B.; Boldyrev, A. I. *J. Phys. Chem. A* **2007**, *111*, 12864–12866.

(4) Tsipis, A. C.; Kefalidis, C. E.; Tsipis, C. A. *J. Am. Chem. Soc.* **2008**, *130*, 9144–9155.

(5) (a) Kuznetsov, A. E.; Boldyrev, A. I.; Li, X.; Wang, L.-S. *J. Am. Chem. Soc.* **2001**, *123*, 8825–8831. (b) Kuznetsov, A. E.; Corbett, J. D.; Wang, L.-S.; Boldyrev, A. I. *Angew. Chem., Int. Ed. Engl.* **2001**, *40*, 3369–3372. (c) Zhan, C.-G.; Zheng, F.; Dixon, D. A. *J. Am. Chem. Soc.* **2002**, *124*, 14795–14803. (d) Zhai, H.-J.; Kuznetsov, A. E.; Boldyrev, A.; Wang, L.-S. *ChemPhysChem* **2004**, *5*, 1885–1891. (e) Liu, Z.-Z.; Tian, W.-Q.; Feng, J.-K.; Zhang, G.; Li, W.-Q. *J. Phys. Chem. A* **2005**, *109*, 5645–5655. (f) Chi, X. X.; Chen, X. J.; Yuan, Z. S. *J. Mol. Struct. (Theochem)* **2005**, *732*, 149–153.

(6) (a) Lin, Y. C.; Jusélius, J.; Sundholm, D.; Gauss, J. *J. Chem. Phys.* **2005**, *122*, 214308. (b) Islas, R.; Heine, T.; Merino, G. *J. Chem. Theory Comput.* **2007**, *3*, 775–781. (c) Kuznetsov, A. E.; Birch, K. A.; Boldyrev, A. I.; Zhai, H.-J.; Wang, L.-S. *Science* **2003**, *300*, 622–625. (d) Ugrinov, A.; Sen, A.; Reber, A. C.; Qian, M.; Khanna, S. N. *J. Am. Chem. Soc.* **2008**, *130*, 782–783.

(7) (a) Kruszewski, J.; Krygowski, T. M. *Tetrahedron Lett.* **1972**, *13*, 3839–3842. (b) Krygowski, T. M. *J. Chem. Inf. Comp. Sci.* **1993**, *33*, 70–78.

(8) Matito, E.; Duran, M.; Solà M. *J. Chem. Phys.* **2005**, *122*, 014109. Erratum: *J. Chem. Phys.* **2006**, *125*, 059901.

(9) Cyrański, M. K. *Chem. Rev.* **2005**, *105*, 3773–3811.

(10) Boldyrev, A. I.; Kuznetsov, A. E. *Inorg. Chem.* **2002**, *41*, 532–537.

(11) (a) Hückel, E. *Z. Physik* **1931**, *70*, 204–286. (b) Hückel, E. *Z. Physik* **1931**, *72*, 310–337. (c) Hückel, E. *Z. Physik* **1932**, *76*, 628–648. (d) Hückel, E. *Z. Elektrochem.* **1937**, *43*, 752–788, and 827−849.

(12) Schleyer, P. v. R.; Maerker, C.; Dransfeld, A.; Jiao, H.; van Eikema Hommes, N. J. R. *J. Am. Chem. Soc.* **1996**, *118*, 6317–6318.

(13) (a) Giambiagi, M.; de Giambiagi, M. S.; dos Santos, C. D.; de Figueiredo, A. P. *Phys. Chem. Chem. Phys.* **2000**, *2*, 3381–3392. (b) Bultinck, P.; Ponec, R.; Van Damme, S. *J. Phys. Org. Chem.* **2005**, *18*, 706–718. (c) Bultinck, P.; Rafat, M.; Ponec, R.; van Gheluwe, B.; Carbó-Dorca, R.; Popelier, P. *J. Phys. Chem. A* **2006**, *110*, 7642–7648.

(14) (a) Mandado, M.; Krishtal, A.; Van Alsenoy, C.; Bultinck, P.; Hermida-Ramón, J. M. *J. Phys. Chem. A* **2007**, *111*, 11885–11893. (b) Roy, D. R.; Bultinck, P.; Subramanian, V.; Chattaraj, P. K. *J. Mol. Struct. (Theochem)* **2008**, *854*, 35–39. (c) Jiménez-Halla, J. O. C.; Matito, E.; Blancafort, L.; Robles, J.; Solà, M. *J. Comput. Chem.* **2009**, *30*, 2764–2776.

(15) (a) Havenith, R. W. A.; Fowler, P. W.; Steiner, E.; Shetty, S.; Kanhere, D.; Pal, S. *Phys. Chem. Chem. Phys.* **2004**, *6*, 285–288. (b) Jung, Y.; Heine, T.; Schleyer, P. v. R.; Head-Gordon, M. *J. Am. Chem. Soc.* **2004**, *126*, 3132–3138.

(16) Fowler, P. W.; Havenith, R. W. A.; Steiner, E. *Chem. Phys. Lett.* **2001**, *342*, 85–90.

(17) Havenith, R. W. A.; van Lenthe, J. H. *Chem. Phys. Lett.* **2004**, *385*, 198–201.

(18) Aihara, J.; Kanno, H.; Ishida, T. *J. Am. Chem. Soc.* **2005**, *127*, 13324–13330.

(19) Corminboeuf, C.; Heine, T.; Weber, J. *Phys. Chem. Chem. Phys.* **2003**, *5*, 246–251.

(20) (a) Chen, Z.; Corminboeuf, C.; Heine, T.; Bohmann, J.; Schleyer, P. v. R. *J. Am. Chem. Soc.* **2003**, *125*, 13930–13931. (b) Wannere, C. S.; Corminboeuf, C.; Wang, Z.-X.; Wodrich, M. D.; King, R. B.; Schleyer, P. v. R. *J. Am. Chem. Soc.* **2005**, *127*, 5701–5705.

(21) Jiménez-Halla, J. O. C.; Matito, E.; Robles, J.; Solà, M. *J. Organomet. Chem.* **2006**, *691*, 4359–4366.

(22) Tsipis, A. C. *Phys. Chem. Chem. Phys.* **2009**, *11*, 8244–8261.

(23) Fowler, P. W.; Havenith, R. W. A.; Steiner, E. *Chem. Phys. Lett.* **2002**, *359*, 530–536.

(24) Feixas, F.; Matito, E.; Poater, J.; Solà, M. *J. Comput. Chem.* **2008**, *29*, 1543–1554.

(25) (a) Jusélius, J.; Straka, M.; Sundholm, D. *J. Phys. Chem. A* **2001**, *105*, 9939–9944. (b) Santos, J. C.; Tiznado, W.; Contreras, R.; Fuentealba, P. *J. Chem. Phys.* **2004**, *120*, 1670–1673. (c) Chatarraj, P. K.; Roy, D. R.; Elango, M.; Subramanian, V. *J. Phys. Chem. A* **2005**, *109*, 9590–9597.

(26) (a) Twamley, B.; Power, P. P. *Angew. Chem., Int. Ed. Engl.* **2000**, *39*, 3500–3503. (b) Kuznetsov, A. E.; Boldyrev, A. I.; Li, X.; Wang, L.-S. *J. Am. Chem. Soc.* **2001**, *123*, 8825–8831.

(27) (a) Li, X.; Zhang, H.-F.; Wang, L.-S.; Kuznetsov, A. E.; Cannon, N. A.; Boldyrev, A. I. *Angew. Chem., Int. Ed. Engl.* **2001**, *40*, 1867–1870. (b) Yang, L.-M.; Wang, J.; Ding, Y.-H.; Sun, C.-C. *Organometallics* **2007**, *26*, 4449–4455.

(28) (a) Seal, P. *J. Mol. Struct. (Theochem)* **2009**, *893*, 31–36. (b) Nigam, S.; Majumder, C.; Kulshreshtha, S. K. *J. Mol. Struct. (Theochem)* **2005**, *755*, 187–194.

(29) Chi, X. X.; Li, X. H.; Chen, X. J.; Yuang, Z. S. *J. Mol. Struct. (Theochem)* **2004**, *677*, 21–27.

(30) (a) Jin, Q.; Jin, B.; Xu, W. G.; Zhu, W. *J. Mol. Struct. (Theochem)* **2005**, *713*, 113–117. (b) Kraus, F.; Korber, N. *Chem.—Eur. J.* **2005**, *11*, 5945–5959. (c) De Proft, F.; Fowler, P. W.; Havenith, R. W. A.; Schleyer, P. v. R.; Van Lier, G.; Geerlings, P. *Chem.—Eur. J.* **2004**, *10*, 940–950.

(31) Frisch, M. J. ; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery Jr., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision C.01; Gaussian, Inc.: Pittsburgh, PA, 2003.

(32) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623–11627.

(33) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(34) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785–789.

(35) (a) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650–654. (b) Frisch, M. J.; Pople, J. A.; Binkley, J. S. *J. Chem. Phys.* **1984**, *80*, 3265–3269.

(36) Lambrecht, D. S.; Fleig, T.; Sommerfeld, T. *J. Phys. Chem. A* **2008**, *112*, 2855–2862.

(37) Zubarev, D. Y.; Boldyrev, A. I. *J. Phys. Chem. A* **2008**, *112*, 7984–7885.

(38) Lambrecht, D. S.; Fleig, T.; Sommerfeld, T. *J. Phys. Chem. A* **2008**, *112*, 7986–7986.

(39) (a) Wolinski, K.; Hilton, J. F.; Pulay, P. *J. Am. Chem. Soc.* **1990**, *112*, 8251–8260. (b) Cheeseman, J. R.; Trucks, G. W.; Keith, T. A.; Frisch, M. J. *J. Chem. Phys.* **1996**, *104*, 5497–5509.

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

(41) (a) Cossio, F. P.; Morao, I.; Jiao, H. J.; Schleyer, P. v. R. *J. Am. Chem. Soc.* **1999**, *121*, 6737–6746. (b) Morao, I.; Lecea, B.; Cossio, F. P. *J. Org. Chem.* **1997**, *62*, 7033–7036.

(42) Schleyer, P. v. R.; Manoharan, M.; Wang, Z. X.; Kiran, B.; Jiao, H. J.; Puchta, R.; van Eikema Hommes, N. J. R. *Org. Lett.* **2001**, *3*, 2465–2468.

(43) Corminboeuf, C.; Heine, T.; Seifert, G.; Schleyer, P. v. R.; Weber, J. *Phys. Chem. Chem. Phys.* **2004**, *6*, 273–276.

(44) (a) Glendening, E. D.; Badenhoop, J. K.; Reed, A. E.; Carpenter, J. E.; Bohmann, J. A.; Morales, C. M.; Weinhold, F. *NBO 5.0 Program*; Theoretical Chemistry Institute, University of Wisconsin, Madison, WI, 2001; (b) Reed, A. E.; Curtiss, L. A.; Weinhold, F. *Chem. Rev.* **1988**, *88*, 899–926.

(45) Cioslowski, J.; Matito, E.; Solà, M. *J. Phys. Chem. A* **2007**, *111*, 6521–6525.

(46) (a) Mandado, M.; González-Moa, M. J.; Mosquera, R. A. *J. Comput. Chem.* **2007**, *28*, 127–136. (b) Mandado, M.; González-Moa, M. J.; Mosquera, R. A. *ChemPhysChem* **2007**, *8*, 696–702.

(47) (a) Matito, E.; Solà, M.; Salvador, P.; Duran, M. *Faraday Discuss.* **2007**, *135*, 325–345. (b) Matito, E.; Poater, J.; Solà, M.; Duran, M.; Salvador, P. *J. Phys. Chem. A* **2005**, *109*, 9904–9910. (c) Ponec, R.; Cooper, D. *J. Mol. Struct. (Theochem)* **2005**, *727*, 133–138.

(48) (a) Bader, R. F. W. *Acc. Chem. Res.* **1985**, *18*, 9–15. (b) Bader, R. F. W. *Chem. Rev.* **1991**, *91*, 893–928.

(49) Biegler-König, F. W.; Bader, R. F. W.; Tang, T.-H. *J. Comput. Chem.* **1982**, *3*, 317–328. (http://www.chemistry.mcmaster.ca/aimpac/).

(50) Matito, E. *ESI-3D: Electron Sharing Indexes Program for 3D Molecular Space Partitioning*; Institute of Computational

Chemistry, University of Girona: Catalonia, Spain, 2006; http://iqc.udg.es/∼eduard/ESI.

(51) The numerical accuracy of the QTAIM calculations has been assessed using two criteria: (i) The integration of the Laplacian of the electron density ($\nabla^2\rho(\mathbf{r})$) within an atomic basin must be close to zero; and (ii) The number of electrons in a molecule must be equal to the sum of all the electron populations of the molecule. For all atomic calculations, integrated absolute values of $\nabla^2\rho(\mathbf{r})$ were always less than 0.001 au. For all molecules, errors in the calculated number of electrons were always below 0.01 au. It is important to mention that the default maximum distance from the nucleus used to integrate the atomic region has to be increased when diffuse functions are employed in the presence of metal atoms. In the AIMPAC program, the default integration maximum distance is 9.0 au. However, we have found that this distance should be increased to 12.0 au for the proper integration of Al and Ge atoms, and to 17.0 au for Ga. If this value is not increased, the sum of all electron populations will not be equal to the number of electrons in a molecule. Consequently, these integration distances have to be changed in the input file.

(52) The reason for the differences found is not totally clear to us, although it may be the result of using the default maximum distance (9 au) for the integration of the atomic regions in the AIMPAC program instead of our recommended value of 12 au for Al.

(53) Feixas, F.; Matito, E.; Solà, M.; Poater, J. *J. Phys. Chem. A* **2008**, *112*, 13231–13238.

CT100034P

JCTC Journal of Chemical Theory and Computation

# Not All That Has a Negative NICS Is Aromatic: The Case of the H-Bonded Cyclic Trimer of HF

Rafael Islas,[†] Gerardo Martínez-Guajardo,[†] J. Oscar C. Jiménez-Halla,[†] Miquel Solà,[‡] and Gabriel Merino[*,†]

*Departamento de Química, División de Ciencias Naturales y Exactas, Universidad de Guanajuato, Col. Noria Alta s/n C.P. 36050, Guanajuato, Gto., México, and Institut de Química Computacional and Departament de Química, Universitat de Girona, Campus de Montilivi, 17071 Girona, Catalonia, Spain*

Received February 19, 2010

**Abstract:** In this work, we used the induced magnetic field ($\mathbf{B}^{ind}$) to study the degree of aromaticity of the planar $(HF)_3$ ring. The induced magnetic field analysis shows that the degree of electron delocalization in the hydrogen-bonded cyclic trimer of HF is very low. This result is in agreement with those obtained using GIMIC and is opposite to the Rehaman's suggestion. Our results demonstrate a clear limitation of the NICS index when a strong anisotropy is exhibited and suggest that the NICS values should be used carefully to discuss aromaticity in systems without an important $p_z$-orbital overlap that produces the $\pi$ clouds. In view of the fact that the NICS index is extensively used by computationally and theoretically oriented experimental chemists, this is an important warning.

## Introduction

The IUPAC defines aromaticity as "*The concept of spatial and electronic structure of cyclic molecular systems displaying the effects of cyclic electron delocalization which provide for their enhanced thermodynamic stability (relative to acyclic structural analogues) and tendency to retain the structural type in the course of chemical transformations*".[1] Nevertheless, it is enough to check the last two *Chem. Rev.* issues dedicated to aromaticity[2] and electron delocalization[3] to realize that this definition is still controversial. In general, a quantitative and/or qualitative assessment of the degree of aromaticity is given by the chemical behavior (lower reactivity), structural features (planarity and equal bond length tendencies),[4,5] energy (stability), and spectroscopic properties (UV, proton chemical shifts, and magnetic susceptibility exaltation). Recently several magnetic indices of aromaticity have been introduced and discussed. They include the famous nucleus-independent chemical shift (NICS) and related indexes,[6,7] aromatic ring-current shielding (ARCS),[8] and plotted ring-current densities.[9,10] Particularly, the concept of a ring current, introduced by Pople, has been used widely to interpret the magnetic properties of aromatic molecules.[11,12]

Even though that aromaticity is not well-defined and was originally developed within the organic chemistry scheme, this concept has been extended to inorganic systems. Interestingly, all-metal clusters and inorganic compounds[13−17] have not only the conventional $\pi$-(anti)aromaticity but also the $\sigma$-,[18−22] $\delta$-[23−25] or even $\varphi$-(anti)aromaticity,[26] i.e., a multifold aromaticity.[27−30]

In 2005 Rehaman et al. suggested that the planar hydrogen-bonded cyclic $(HF)_3$ (see Scheme 1) is aromatic.[31] They claimed that the existence of aromaticity in such hydrogen-bond complexes is apparent from the NICS values (in ppm) of −2.94, −1.98, and −1.89 for $(HF)_3$, $(HCl)_3$, and $(HBr)_3$, respectively. Similar results were also reported for water clusters $(H_2O)_n$.[32] Recently, one of us has studied the interplay between aromaticity and hydrogen bonding in 1,3-dihydroxyaryl-2-aldehydes.[33,34] In that case, the quasi-ring partially adopts the role of a typical aromatic ring, favoring phenomena like for proton transfer reactions, contrary to the $(HF)_3$ complex.

Of course, the existence of " hydrogen-bonded aromaticity" involving strong ring currents across hydrogen bonds is

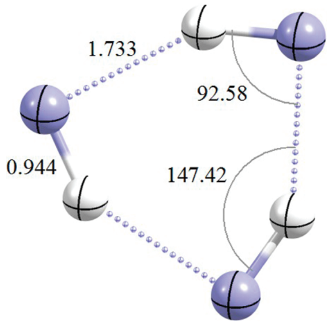
[*] gmerino@quijote.ugto.mx.

[†] Universidad de Guanajuato.

[‡] Universitat de Girona.

10.1021/ct100098c © 2010 American Chemical Society
Published on Web 03/26/2010

*Scheme 1*



very attractive. However, Lin and Sundholm reported the nuclear magnetic shieldings and magnetically induced ring currents for the planar ring-shaped hydrogen fluoride trimer $(HF)_3$,[35] and using the gauge-including magnetically induced current (GIMIC) method[36] showed that, contrary to the Rehaman et al. suggestion,[31] $(HF)_3$ has a very small ring-current susceptibility of 0.37 nA/T. Thus, only a weak net current is passing across the H•••F hydrogen bond.

A magnetic field perpendicular to any plane can induce a current density in and parallel to the selected plane. This current density induces a counter field. In this sense, the induced magnetic fields and induced current densities are complementary to each other. Induced current densities are usually given in one selected plane parallel to the molecular ring, while the induced magnetic field contains information of the overall current density distribution. Having both analyses available is, therefore, advantageous. In other words, the induced magnetic field reveals important information on electron delocalization and, furthermore, of its origin.[37] Recently, we have studied the induced magnetic field of several systems.[38−44] In agreement with current density maps,[10] the response of aromatic, antiaromatic, and nonaromatic examples show different features.

In this work, we used the induced magnetic field ($\mathbf{B}^{ind}$) to study the degree of aromaticity of the planar $(HF)_3$ ring. Similar to the GIMIC results, our calculations show that the mentioned complex is not an aromatic system. Frequently, NICS is related to the strength of the induced ring current. However, in this case, NICS contains contributions from the in-plane components, which can be regarded sometimes as spurious for evaluating aromaticity.

## Computational Details

The geometry optimizations were performed with the B3LYP[45,46] functional, as implemented in the Gaussian 03 program,[47] in conjunction with the def2-TZVPP basis set. At variance with the HF dimer,[48,49] the optimized molecular structure of the trimer is not significantly affected by the basis set superposition error (BSSE) and, therefore, we have not considered the BSSE correction in the HF trimer optimization. The induced magnetic field ($\mathbf{B}^{ind}$) calculations were performed by using the PW91 density functional[50] in conjunction with the IGLO-III basis set. The shielding tensors were computed using the IGLO method.[51] The deMon program was used to compute the molecular orbitals[52] and the deMon-NMR package for the shielding tensors.[53] Induced magnetic fields of the external field applied perpendicularly to the molecular plane were computed in ppm. Assuming an external magnetic field of $|\mathbf{B}^{ext}| = 1.0$ T, the unit of $\mathbf{B}^{ind}$ is 1.0 $\mu$T, which is equivalent to 1.0 ppm of the shielding tensor. In order to render the induced magnetic fields, the molecules were oriented so that the center of mass was located at the origin of the coordinate system; the $z$-axis is parallel to the highest order symmetry axis of the molecule. The external field was applied perpendicular to the $(HF)_3$ plane. VU was employed for the visualization of molecular fields.[54]

## Results and Discussion

Figure 1 depicts the contour lines and isosurfaces of both the $z$-component of the induced magnetic field, $\mathbf{B}^{ind}{}_z$ and the NICS for the $(HF)_3$ complex. It is important to remark that the $\mathbf{B}^{ind}{}_z$ for an external field perpendicular to the ring is equivalent to the $\mathrm{NICS}_{zz}$ index. In a typical aromatic molecule, like benzene, no paratropic contributions to the $\mathbf{B}^{ind}$ are observed inside the ring, and only a strong shielding region to the carbons inside the ring is obtained. In contrast, the hydrogen complex shows a strong but short-range paratropic response inside the ring (Figure 1A). Interestingly, the shielding cones above and below the $(HF)_3$ complex are comparable in shape and intensity to those of a nonaromatic system.[37] Note that each HF moiety is diatropic, but the deshielding regions are further outside (given in red), and there is not an effective overlap of the HF diatropic zones close to the plane ring.

NICS, defined as the negative total isotropic shielding (average shielding), can be computed at any point in space. Isolines and the isosurface of NICS, that is, the effect on the isotropic shielding, caused by external magnetic fields from all directions, are depicted in Figure 1B. The average virtual shielding (3.7 ppm) evaluated at the center of mass corresponds to a NICS index of −3.7, according to the definition proposed by Schleyer et al. (cf. with the value of −2.9 as reported by Rehaman et al.).[31] It must be remarked that the large anisotropies of the shielding tensors of the $(HF)_3$ complex lead to a notable change between both $\mathbf{B}^{ind}{}_z$ and NICS scalar fields: (i) The isotropic shielding is smaller in magnitude than that caused only by an external field in the $z$-direction and (ii) Close to the ring center, the NICS values are negative showing a low aromatic character. Generally, the NICS tensor is strongly anisotropic, i.e., values of $\sigma_{xx}$ and $\sigma_{yy}$ components differ strongly from the $\sigma_{zz}$ contribution; however, in most cases, like in benzene, the absolute value of the $(\sigma_{xx} + \sigma_{yy})$ term is smaller than that of the $\sigma_{zz}$ one, therefore, the $\mathbf{B}^{ind}{}_z$ and NICS have the same sign. In this case, at the ring center the $\sigma_{xx} = \sigma_{yy} = 11.5$ ppm and the $\sigma_{zz} = -13.3$ ppm, giving a big difference between both aromaticity indexes.

Formally, the pure $p_z$ lone pairs of the fluorine atoms become three $\pi$ orbitals of the planar $(HF)_3$ complex, so the system could be considered a Hückel aromatic complex. However, there is not a remarkable overlap of $p_z$ orbitals that produces the $\pi$ clouds and, thus, there is not an important electron delocalization in that complex. As our method allows the separation into core $\sigma$ and $\pi$ orbitals, we can discuss the

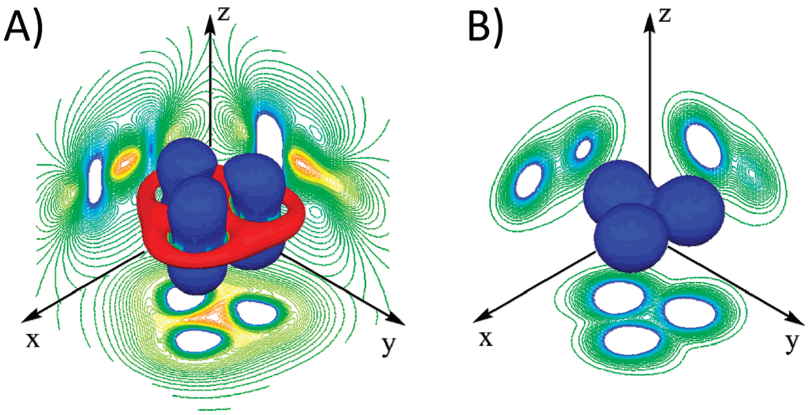


***Figure 1.*** (A) Isosurfaces of the *z*-component of the induced magnetic field $\mathbf{B}^{ind}_z$ and (B) NICS. $|\mathbf{B}^{ind}_z|$ and $|NICS(\mathbf{r})| = 4.0\ \mu T$, and $\mathbf{B}^{ext} = 1.0$ T perpendicular to the molecular plane. Blue and red colors indicate shielding and deshielding areas, respectively.

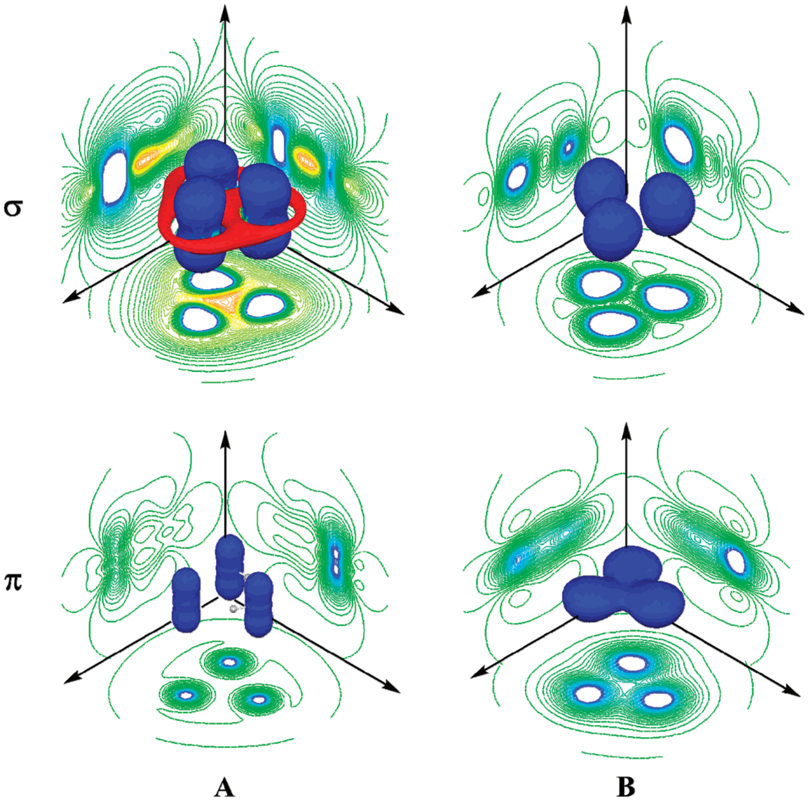


***Figure 2.*** Induced magnetic field and NICS of the $(HF)_3$ complex. (A) The $\mathbf{B}^{ind}_z$, shielding (diatropic, blue) or enforcing (paratropic, red) the external field, shown in the same planes as for the field lines. (B) Contour lines of NICS($\mathbf{r}$) (equivalent to the negative shielding density) are shown in the same planes as the field lines.

role of each orbital contribution as shown in Figure 2. The core electrons do not contribute to $\mathbf{B}^{ind}$, except in the very close vicinity of nuclei. The $\sigma$ orbital contributions to $\mathbf{B}^{ind}_z$ have quite a similar shape and magnitude to the total $\mathbf{B}^{ind}_z$. As $\sigma$ electrons are strongly localized, their local diamagnetic contributions generate a short-range response and a paratropic (deshielding) region at the ring center (Figure 2A). In the same way, the $\pi$ electrons are quite localized at the fluorine atoms, and only a short-range diamagnetic response is observed. In contrast, the $NICS_\sigma$ still includes the current effects induced by magnetic fields parallel to the ring, which are not negligible and have an impact in the shape of the shielding cones. The $\sigma$ component is responsible for the negative value observed at the ring center. Lazzeretti pointed out that serious interpretation errors can be done by confusing the averaged trace and the out-of-plane component.[10] Our results are in agreement with that comment.

Finally, let us estimate the contribution of each H−F fragment to both the induced magnetic field and NICS. Of course, this approximation is only valid for weak interactions. For instance, Poater and co-workers showed that the apparent increase of local aromaticity in superimposed aromatic rings indicated by NICS is not real, but rather the result of the magnetic field generated in one ring due to the electron

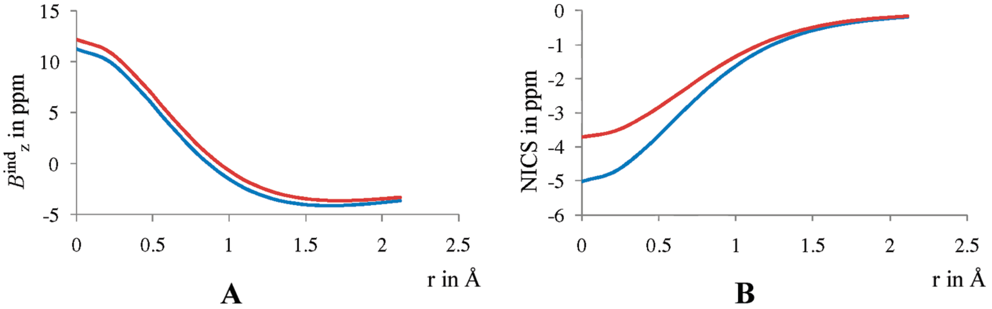


***Figure 3.*** (A) $\mathbf{B}^{ind}_z$ and (B) NICS profiles along the $z$-axis of the planar $(HF)_3$ trimer. The blue and red lines show the profiles of the noninteracting and bonded complexes, respectively.

current density of the other ring placed above it.[55] In our case, Figure 3 shows the profiles along the $z$-axis for both scalar fields ($\mathbf{B}^{ind}_z$ and NICS); $r = 0$ corresponds to the ring center. Interestingly, at $r = 0$ the contribution of each fragment to $\mathbf{B}^{ind}_z$ is 3.75 ppm, i.e., the sum of three fragments is 1.0 ppm lower than in the complex (12.2 ppm). In this sense, our results provide evidence that the noninteracting trimer is more diatropic (aromatic) than the planar bonded $(HF)_3$ complex. A similar conclusion can be emerged from the NICS analysis.

## Conclusion

We have shown that the electron delocalization degree in the planar hydrogen-bonded HF cyclic trimer is very low. This result is in agreement with that obtained using GIMIC[36] and is opposite to the Rehaman et al. suggestion.[31] Our results show a clear limitation of the NICS index when a strong anisotropy is exhibited and suggest that the NICS values should be used carefully to discuss aromaticity in systems without an important overlap of $p_z$ orbitals that produces the $\pi$ clouds. In view of the fact that the NICS index is extensively used by computationally and theoretically oriented experimental chemists, this is an important warning. Our results are also in line with the Lazzeretti comments,[10] who mentioned that the analyses based on average values imply a loss of information, and thus criteria for diatropicity and aromaticity should be established in terms of the out-of-plane component of magnetic tensors. Our conclusion to be cautious regarding an interpretation of the NICS index is also supported by the work of Pierrefixe et al.[5] They showed that the symmetric geometry in benzene-type aromatic species (which is originating the small highest-occupied and lowest-unoccupied molecular orbitals, HOMO−LUMO, gap of the $\pi$ system and thus contributes to the ring current) is caused by the tendency of the $\sigma$ system, not the $\pi$ system, which is still often (erroneously) held responsible.

**Acknowledgment.** The work in Guanajuato was supported by the Air Force Research Laboratory *Latin American Initiative Project*, Conacyt (Grant 57892) and DAIP-UGTO. R.I. and G.M.G. gratefully acknowledge Conacyt for Ph.D. fellowships. M.S. thanks the Spanish MICINN (project CTQ2008-03077/BQU) and the Catalan DIUE (project 2009SGR637 and ICREA Academia 2010 Prize) for their financial support.

## References

(1) Minkin, V. I. *Pure Appl. Chem.* **1999**, *71*, 1919.

(2) Schleyer, P. v. R. *Chem. Rev.* **2001**, *101*, 1115.

(3) Schleyer, P. v. R. *Chem. Rev.* **2005**, *105*, 3433.

(4) Pierrefixe, S. C. A. H.; Bickelhaupt, F. M. *Chem.—Eur. J.* **2007**, *13*, 6321.

(5) Pierrefixe, S. C. A. H.; Bickelhaupt, F. M. *J. Phys. Chem. A* **2008**, *112*, 12816.

(6) Schleyer, P. v. R.; Maerker, C.; Dransfeld, A.; Jiao, H. J.; Hommes, N. J. R. v. E. *J. Am. Chem. Soc.* **1996**, *118*, 6317.

(7) Chen, Z. F.; Wannere, C. S.; Corminboeuf, C.; Puchta, R.; Schleyer, P. v. R. *Chem. Rev.* **2005**, *105*, 3842.

(8) Jusélius, J.; Sundholm, D. *Phys. Chem. Chem. Phys.* **1999**, *1*, 3429.

(9) Gomes, J.; Mallion, R. B. *Chem. Rev.* **2001**, *101*, 1349.

(10) Lazzeretti, P. *Prog. Nucl. Magn. Reson. Spectrosc.* **2000**, *36*, 1.

(11) Pople, J. A. *Trans. Faraday Soc.* **1953**, *49*, 1375.

(12) Pople, J. A. *J. Chem. Phys.* **1958**, *24*, 1111.

(13) Li, X.; Kuznetsov, A. E.; Zhang, H. F.; Boldyrev, A. I.; Wang, L.-S. *Science* **2001**, *291*, 859.

(14) Kuznetsov, A. E.; Birch, K. A.; Boldyrev, A. I.; Li, X.; Zhai, H.-J.; Wang, L.-S. *Science* **2003**, *300*, 622.

(15) Tsipis, C. A. *Coord. Chem. Rev.* **2005**, *249*, 2740.

(16) Zubarev, D. Y.; Averkiev, B. B.; Zhai, H.-J.; Wang, L.-S.; Boldyrev, A. I. *Phys. Chem. Chem. Phys.* **2008**, *10*, 257.

(17) Jimenez-Halla, J. O. C.; Matito, E.; Robles, J.; Solà, M. *J. Organomet. Chem.* **2006**, *691*, 4359.

(18) Alexandrova, A. N.; Boldyrev, A. I. *J. Phys. Chem. A* **2003**, *107*, 554.

(19) Pelloni, S.; Lazzeretti, P.; Zanasi, R. *J. Phys. Chem. A* **2007**, *111*, 8163.

(20) Wu, W.; Ma, B.; Wu, J. I. C.; Schleyer, P. v. R.; Mo, Y. R. *Chem.—Eur. J.* **2009**, *15*, 9730.

(21) Jimenez-Halla, J. O. C.; Matito, E.; Blancafort, L.; Robles, J.; Solà, M. *J. Comput. Chem.* **2009**, *30*, 2764.

(22) Giri, S.; Roy, D. R.; Duley, S.; Chakraborty, A.; Parthasarathi, R.; Elango, M.; Vijayaraj, R.; Subramanian, V.; Islas, R.; Merino, G.; Chattaraj, P. K. *J. Comput. Chem.* **2010**, DOI: 10.1002/jcc.21452

(23) Huang, X.; Zhai, H. J.; Kiran, B.; Wang, L. S. *Angew. Chem., Int. Ed.* **2005**, *44*, 7251.

(24) Averkiev, B. B.; Boldyrev, A. I. *J. Phys. Chem. A* **2007**, *111*, 12864.

(25) Zhai, H.-J.; Averkiev, B. B.; Zubarev, D. Y.; Wang, L.-S.; Boldyrev, A. I. *Angew. Chem., Int. Ed.* **2007**, *46*, 4277.

(26) Tsipis, A. C.; Kefalidis, C. E.; Tsipis, C. A. *J. Am. Chem. Soc.* **2008**, *130*, 9144.

(27) Kuznetsov, A. E.; Boldyrev, A. I.; Li, X.; Wang, L.-S. *J. Am. Chem. Soc.* **2001**, *123*, 8825.

(28) Kuznetsov, A. E.; Corbett, J. D.; Wang, L.-S.; Boldyrev, A. I. *Angew. Chem., Int. Ed.* **2001**, *40*, 3369.

(29) Boldyrev, A. I.; Wang, L.-S. *Chem. Rev.* **2005**, *105*, 3716.

(30) Zhan, C. G.; Zheng, F.; Dixon, D. A. *J. Am. Chem. Soc.* **2002**, *124*, 14795.

(31) Rehaman, A.; Datta, A.; Mallajosyula, S. S.; Pati, S. K. *J. Chem. Theory Comput.* **2006**, *2*, 30.

(32) Datta, A.; Pati, S. K. *Int. J. Quantum Chem.* **2006**, *106*, 1697.

(33) Palusiak, M.; Simon, S.; Solà, M. *Chem. Phys.* **2007**, *342*, 43.

(34) Palusiak, M.; Simon, S.; Solà, M. *J. Org. Chem.* **2009**, *74*, 2059.

(35) Lin, Y. C.; Sundholm, D.; Jusélius, J. *J. Chem. Theory Comput.* **2006**, *2*, 761.

(36) Jusélius, J.; Sundholm, D.; Gauss, J. *J. Chem. Phys.* **2004**, *121*, 3952.

(37) Merino, G.; Heine, T.; Seifert, G. *Chem.—Eur. J.* **2004**, *10*, 4367.

(38) Islas, R.; Chamorro, E.; Robles, J.; Heine, T.; Santos, J. C.; Merino, G. *Struct. Chem.* **2007**, *18*, 833.

(39) Islas, R.; Heine, T.; Ito, K.; Schleyer, P. v. R.; Merino, G. *J. Am. Chem. Soc.* **2007**, *129*, 14767.

(40) Islas, R.; Heine, T.; Merino, G. *J. Chem. Theory Comput.* **2007**, *3*, 775.

(41) Merino, G.; Mendez-Rojas, M. A.; Beltraan, H. I.; Corminboeuf, C.; Heine, T.; Vela, A. *J. Am. Chem. Soc.* **2004**, *126*, 16160.

(42) Perez-Peralta, N.; Heine, T.; Barthel, R.; Seifert, G.; Vela, A.; Mendez-Rojas, M. A.; Merino, G. *Org. Lett.* **2005**, *7*, 1509.

(43) Perez-Peralta, N.; Sanchez, M.; Martin-Polo, J.; Islas, R.; Vela, A.; Merino, G. *J. Org. Chem.* **2008**, *73*, 7037.

(44) Tiznado, W.; Perez-Peralta, N.; Islas, R.; Toro-Labbe, A.; Ugalde, J. M.; Merino, G. *J. Am. Chem. Soc.* **2009**, *131*, 9426.

(45) Becke, A. D. *Phys. Rev. A: At., Mol., Opt. Phys.* **1988**, *38*, 3098.

(46) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785.

(47) 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*; Gaussian, Inc.: Wallington, CT, 2003.

(48) Salvador, P.; Paizs, B.; Duran, M.; Suhai, S. *J. Comput. Chem.* **2001**, *22*, 765.

(49) Salvador, P.; Paizs, B.; Duran, M.; Suhai, S. *J. Comput. Chem.* **2006**, *27*, 1505.

(50) Perdew, J. P. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1986**, *33*, 8822.

(51) Kutzelnigg, W. *Isr. J. Chem.* **1980**, *19*, 193.

(52) Köster, A. M.; Calaminici, P.; Flores-Moreno, R.; Geudtner, G.; Goursot, A.; Heine, T.; Janetzko, F.; Patchkovskii, S.; Reveles, J. U.; Vela, A.; Salahub, D. R. *deMon2k*; The deMon Developers Community: Cinvestav, Mexico, 2008.

(53) Malkin, V. G.; Malkina, O. L.; Salahub, D. R. *Chem. Phys. Lett.* **1993**, *204*, 80.

(54) Ozell, B.; Camarero, R.; Garon, A.; Guibault, F. *Finite Elem. Anal. Des.* **1995**, *19*, 295.

(55) Poater, J.; Bofill, J. M.; Alemany, P.; Solà, M. *J. Org. Chem.* **2006**, *71*, 1700.

CT100098C

JCTC Journal of Chemical Theory and Computation

# Constrained Broyden Minimization Combined with the Dimer Method for Locating Transition State of Complex Reactions

Cheng Shang and Zhi-Pan Liu*

*Shanghai Key Laboratory of Molecular Catalysis and Innovative Materials, Department of Chemistry, MOE Key Laboratory for Computational Physical Sciences, Fudan University, Shanghai 200433, China*

Received September 28, 2009

**Abstract:** To determine transition state (TS) and, thus, to predict chemical activity has been a challenging topic in theoretical simulations of chemical reactions. In particular, with the difficulty to compute the second derivative of energy (Hessian) in modern quantum mechanics packages with a non-Gaussian basis set, the location usually involves a high demand in computational power and lacks stability in the algorithm, especially for complex reaction systems with many degrees of freedom. Here, an efficient TS searching method is developed by combining the constrained Broyden minimization algorithm with the dimer method that was first proposed by Henkelman and Jónsson. In the new method, the rotation of the dimer needs only one energy and gradient calculation for determining a rotation angle; the translation of the dimer is continually carried out until a termination criterion is met, and the translational force parallel to the dimer direction is damped to optimize the searching trajectory. Based on our results of the Baker reaction system and of a heterogeneous catalytic reaction, our method is shown to increase the efficiency significantly and is also more stable in finding TSs.

## 1. Introduction

Transition-state theory (TST) plays a central role in chemical kinetics. To determine TS and, thus, to predict chemical activity based on TST is a major theme in modern theoretical simulation of chemical reactions. The algorithms for locating TS can be generally divided into two classes, namely, (i) chain-of-states and (ii) surface-walking methods. The former class simultaneously optimizes a few connected images on the potential energy surface (PES) to identify the minimum energy path (MEP). The representative methods include the nudge elastic band (NEB),[1–5] the doubly nudge elastic band (DNEB),[6–9] and the string methods.[10,11] The later class optimizes only one structural image on the PES by using the local information, such as the gradient (force) or the second derivative (Hessian) of PES. As a result, these methods are much less demanding in computational power. Belonging to this category are the methods, such as the partitioned rational function optimizer (P-RFO),[12–15] the hybrid eigenvector following,[16,17] the dimer,[18–21] and the bond-length constrained minimization methods.[22,23]

Among all the methods in searching for TS, the Hessian involved methods, such as the P-RFO approach, are perhaps the most efficient when the (analytic) Hessian is cheaply available.[15] By modifying and following the eigenvalue of Hessian, these methods can maximize energy in one degree of freedom, while minimizing energy in all the others. To reduce the computational cost in calculating Hessian, the quasi-Newton-based methods have been utilized to update the Hessian, such as the Powell symmetric Broyden (PSB), the symmetric rank 1 (Murtagh−Sargent, MS)[24–28] and the hybrid approach developed by Bofill.[29–34] In practice, a constraint on the step length is often implemented to deal with the overstepping problem.[35] A comprehensive survey for methods in this category has been reviewed by Schlegel.[36]

Different from the P-RFO approach, the dimer method, proposed by Henkelman and Jónsson[18] initially and developed by several other groups[19–21] later, can locate the TS without the need of Hessian. This is particularly advanta-

* Email address: zpliu@fudan.edu.cn.

10.1021/ct9005147 © 2010 American Chemical Society
Published on Web 03/10/2010

geous for the cases: (i) when the Hessian is not cheaply available, as in modern quantum mechanics packages with non-Gaussian basis sets and (ii) where a large reaction system with many degrees of freedom is interested.[20,21] The dimer method involves two structural images (defined as a dimer) on the PES, the linkage between which creates a unit vector $\hat{\mathbf{N}}$ separated by a prefixed distance $2\Delta R$ (e.g., $\Delta R = 0.005$ Å). The whole algorithm of the dimer method is constituted by two independent parts, namely rotation and translation. The dimer first rotates to identify the local curvature ($C$), which is equivalent to determine a relevant normal mode of Hessian numerically. The curvature $C$ is calculated using eq 1, where the vectors $\mathbf{F}_2$ and $\mathbf{F}_1$ are the forces acting on each image of the dimer. Then the dimer translates toward the TS. The force for the dimer rotation ($\Delta\mathbf{F}^{\perp}$) and the translation ($\mathbf{F}_{\text{tran}}$) are described by eqs 2 and 3, respectively. In eq 3, $\mathbf{F}_0$ is the total force acting on the middle point of the dimer. The parallel force $\mathbf{F}^{\|}$ ($\mathbf{F}^{\|} \equiv (\mathbf{F}_0 \bullet \hat{\mathbf{N}})\cdot\hat{\mathbf{N}}$) is defined as the force component parallel to the dimer direction $\hat{\mathbf{N}}$, and the vertical force $\mathbf{F}^{\perp}$ is defined by $\mathbf{F}^{\perp} \equiv \mathbf{F}_0 - \mathbf{F}^{\|}$.

$$C = \frac{(\mathbf{F}_1 - \mathbf{F}_2)\bullet\hat{\mathbf{N}}}{2\Delta R} \tag{1}$$

$$\Delta\mathbf{F}^{\perp} = (\mathbf{F}_1 - \mathbf{F}_2) - [(\mathbf{F}_1 - \mathbf{F}_2)\bullet\hat{\mathbf{N}}]\cdot\hat{\mathbf{N}} \tag{2}$$

$$\mathbf{F}_{\text{tran}} = \begin{cases} -\mathbf{F}^{\|} = -(\mathbf{F}_0\bullet\hat{\mathbf{N}})\cdot\hat{\mathbf{N}} & (C > 0) \\ \mathbf{F}^{\perp} - \mathbf{F}^{\|} = \mathbf{F}_0 - 2(\mathbf{F}_0\bullet\hat{\mathbf{N}})\cdot\hat{\mathbf{N}} & (C < 0) \end{cases} \tag{3}$$

In the first version of the dimer method, four energy and gradient calculations are required to determine a rotation angle.[18] To reach a converged curvature, the dimer usually needs to rotate 5−10 times, each with a determined rotation angle. The rotation of the dimer is, therefore, the most time demanding part in TS searching. To improve the efficiency, Olsen et al.[19] suggested that the force on the image 2($\mathbf{F}_2$) can be approximated as $\mathbf{F}_2 = 2\mathbf{F}_0 - \mathbf{F}_1$, which can save two energy and gradient calculations in the determination of a rotation angle. Heyden et al. reported a new method to calculate the rotation angle by expanding the rotation angle using the Fourier series.[20] Based on Heyden's approach, Kästner and Sherwood suggested that the rotational force could be extrapolated to save one more energy and gradient calculation, which, however, may hamper the convergence of the rotation.[21] They also reported that by replacing the conjugate gradient (CG) algorithm with the Broyden−Fletcher−Goldfarb−Shanno (L-BFGS) algorithm in the rotation and translation optimizations, the efficiency of the dimer method could be improved.

Recently, we developed a constrained Broyden minimization (CBM) method to locate TS,[23] where the distance of a chemical bond is fixed and autoupdated by the quasi-Newton Broyden method during the TS searching. The method also does not require Hessian as input. Compared to the dimer method, we found that the CBM method eliminates completely the rotation steps as the reaction coordinate is predefined, simply as a bonding pair, and that the CBM carries out the geometry relaxation with multiple Broyden steps at each fixed bond distance, while the dimer method has only one move (e.g., ≤0.2 Å) at each translation step. These two features make the CBM method efficient in finding simple bond-making/breaking reactions on complex substrates, for example, those often involved in heterogeneous catalysis.[23] However, the CBM method has its own limitation due to the lack of the curvature information; it meets great difficulties in finding complex TS where the reaction coordinate is not so intuitive.

In this work, we aim to develop a better approach by combining the advantages of the dimer and CBM methods to improve the efficiency and the stability in searching for the TS of complex reactions. Indeed, we find that it is possible to integrate the Broyden algorithm in both the rotation and the translation parts of the dimer method. Specifically, our new approach can achieve the following: (i) only one energy and gradient calculation for determining a rotation angle in the dimer rotation; (ii) multiple optimization steps in the dimer translation, similar to the CBM method; and (iii) optimized TS searching trajectory with enhanced stability. By applying to the example reactions, we show that the new algorithm is much more efficient and stable than that of the existing dimer method.

## 2. Methods

Following the terminology of the dimer method, we also describe our algorithm in two sections, namely, the rotation and the translation.

**2.1. Rotation.** The rotation of a dimer can be considered as moving image one ($\mathbf{R}_1$) on a spherical surface, and the center of the sphere is the middle point of the dimer ($\mathbf{R}_0$) with the radius being half of the length of the dimer ($\Delta\mathbf{R}$). The rotation direction was suggested as eq 4, according to Fourier series expansion.[20] To determine the rotation angle $\phi_{\min}$, one needs first make a trial rotation of angle $\phi_1$ to obtain the value of $|\Delta\mathbf{F}^{\perp}_{\phi_1}|$. In total, two energy and gradient calculations ($\Delta\mathbf{F}\perp$ and $\Delta\mathbf{F}^{\perp}_{\phi_1}$) are thus essential for one rotation (rotate a $\phi_{\min}$). Equation 4 may be modified slightly by utilizing the projected $|\Delta\mathbf{F}^{\perp}_{\phi_1}|$ expressed as $(\Delta\mathbf{F}^{\perp}_{\phi_1} \bullet \Delta\mathbf{F}^{\perp})/|\Delta\mathbf{F}^{\perp}|$, instead of $|\Delta\mathbf{F}^{\perp}_{\phi_1}|$.[37] The rotation of the dimer is terminated only if the $\Delta\mathbf{F}\perp$ is below a preset criterion (e.g., rms force ($|\mathbf{F}|$) < 0.05 eV/Å), which typically requires more than five rotations (i.e., ten energy and gradient calculations).

$$\varphi_{\min} = \frac{1}{2}\arctan\left(\frac{\sin(2\varphi_1)\cdot|\Delta\mathbf{F}^{\perp}|}{|\Delta\mathbf{F}^{\perp}|\cdot\cos(2\varphi_1) - |\Delta\mathbf{F}^{\perp}_{\varphi_1}|}\right) \tag{4}$$

We note that by merely minimizing the rotational force $\Delta\mathbf{F}^{\perp}$ in the rotational step, the determined curvature of the dimer may not necessarily be the lowest eigenvalue (normal mode) of Hessian, since $\Delta\mathbf{F}^{\perp}$ is diminished at any eigenvalue of Hessian. Heyden et al.,[20] using the P-RFO method with the computed Hessian, showed that the TS searching by always following the lowest curvature could be unstable, i.e., either converge to the wrong TS or fail to converge. Therefore, an algorithm that can most efficiently reduce the rotational force $\Delta\mathbf{F}^{\perp}$ (i.e., finding the local curvature) is perhaps the most appropriate for the dimer method, provided that a reasonably guessed initial curvature is available.

Following the CBM method, in this work, we utilized the quasi-Newton Broyden method to minimize the rotational force of the dimer. This is addressed in eq 5, where $x$ is chosen as the middle point of the dimer $\mathbf{R}_0$ and $\mathbf{R}_1$, and the residual $R$ is the rotation force $\Delta\mathbf{F}^{\perp}$. It might be mentioned that the Broyden method has been widely used in electronic structure calculations for both charge density mixing and structural optimization.[23,38] The Broyden method[38-41] iteratively updates its Jacobian ($\mathbf{J}$) matrix (approximate Hessian) or the inverse Jacobian ($G$) based on the iteration history, the equations of which have been derived based on the least-squares minimization of an error function, eq 6. The formula for the modified Broyden method as derived by Johnson[41] is summarized in eqs 7−11. The modified Broyden method has no requirement to store the full (3N × 3N) Hessian matrix.[41,42]

$$x_{i+1} = x_i - \mathbf{J}^{-1}R_i \tag{5}$$

$$E = w_0^2||G^{(m+1)} - G^{(m)}|| + \sum_{n=1}^{m} w_n^2\Big||\Delta x^{(n)}\rangle + G^{(m+1)}\Big||\Delta R^{(n)}\rangle|^2 \tag{6}$$

$$G^{(m+1)} = G^{(1)} - \sum_{k=1}^{m} |Z_k^{(m)}\rangle\langle\Delta R^{(k)}| \tag{7}$$

$$|Z_k^{(m)}\rangle = \sum_{n=1}^{m} \beta_{kn}|u^{(n)}\rangle + w_0^2 \sum_{n=1}^{m-1} \beta_{kn}|Z_n^{(m-1)}\rangle \tag{8}$$

$$|u^{(n)}\rangle = G^{(1)}|\Delta R^{(n)}\rangle + |\Delta x^{(n)}\rangle \tag{9}$$

$$\beta_{kn} = (w_0^2 I + a)_{kn}^{-1} \tag{10}$$

$$a_{ij} = w_i w_j |\Delta R^{(j)}\rangle\langle\Delta R^{(i)}| \tag{11}$$

$$\mathbf{R}_1^{\text{new}} = \frac{\mathbf{R}_1' - \mathbf{R}_0}{\Delta R'} \times \Delta R \tag{12}$$

It is noted that because $\Delta\mathbf{F}^{\perp}$ is normal to the dimer $\mathbf{N}$, a direct act of $\Delta\mathbf{F}^{\perp}$ on the dimer will drive the image $\mathbf{R}_1$ away from the spherical surface of the rotation. To restore the dimer length, we utilize eq 12 to constrain $\mathbf{R}_1$ back to the sphere of rotation along the new dimer direction, where $\mathbf{R}_1'$ and $\mathbf{R}_1^{\text{new}}$ are the images before and after the constraint, respectively, and $\Delta R'$ is the length between $\mathbf{R}_1'$ and $\mathbf{R}_0$ ($\Delta R' = |\mathbf{R}_1' - \mathbf{R}_0|$).

**2.2. Translation.** Strictly speaking, the translation direction of the dimer, as guided by the curvature from the rotation, is only meaningful locally, i.e., in a small region defined by $\sim\Delta R$ length on the PES. However, the magnitude of even one translational move has to be very large in practice (e.g., 0.2 Å, 40 times larger than $\Delta R$). This would imply that the traditional framework of the dimer method: one-rotation plus one-translation is not necessarily the safest and the most efficient approach. In principle, it would be desirable to achieve continuous translational moves, i.e., one rotation plus multiple translation, especially with quasi-Newton methods (e.g., BFGS, Broyden) that rely on iteration history to approximate Hessian. With such a framework, the computational cost may be much reduced, since the rotation of the dimer dominates the computational efforts (in doing energy and gradient calculations). The multiple translation steps with an approximate normal mode can, indeed, be used to locate the TS, as already demonstrated in the CBM method in our recent work,[23] where a bond distance is fixed during geometry relaxation. The key challenge is, therefore, to identify a valid criterion for the termination of multiple translation steps.

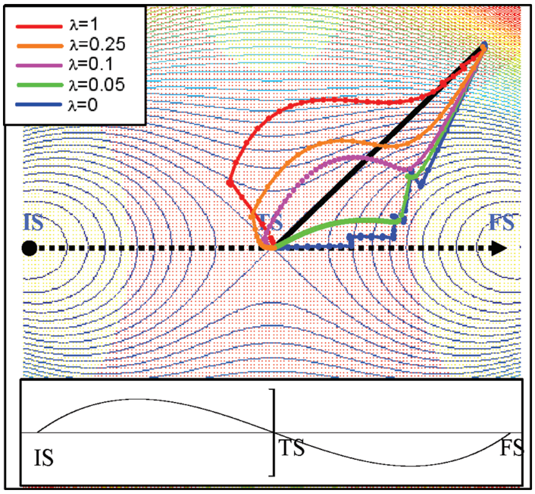


***Figure 1.*** TS-searching trajectories on a 2D PES defined by $E = x^4 + 4x^2y^2 - 2x^2 + 2y^2$. The red region is with one negative curvature (mode), and the inflection point locates at the edge of the red region. The insertion at the bottom of the figure shows how $|\mathbf{F}^{||}|$ varies from the IS to the FS along MEP, as labeled by the dotted arrow. The dotted color curves represent the trajectories with different $\lambda$ values as shown in eq 14.

*2.2.1. Testing in a Two-Dimensional PES.* To find a suitable criterion, we first investigated a simple two-dimensional (2D) PES defined as $E = x^4 + 4x^2y^2 - 2x^2 + 2y^2$, as shown in Figure 1, where the initial state (IS), the TS, and the final state (FS) are labeled. In the figure, the red region contains one negative mode, and the other areas are all positive-curvature regions. The inflection point is located at the edge of the red region. In the insertion, we show that, going from the IS/FS to the TS along MEP, the parallel force $\mathbf{F}^{||}$ has always a maximum in absolute value, which occurs at the inflection point. This means that one should minimize the parallel force at the negative-curvature region but maximize the parallel force at the positive-curvature region. Based on this, we have utilized the following criterion for the termination of the translational move:

$$\begin{cases} |\mathbf{F}_i^{||}| > |\mathbf{F}_{i-1}^{||}| & (C < 0) \\ |\mathbf{F}_i^{||}| < |\mathbf{F}_{i-1}^{||}| & (C > 0) \end{cases} \tag{13}$$

where $\mathbf{F}_i^{||}$ is the $\mathbf{F}^{||}$ of the current step and $\mathbf{F}_{i-1}^{||}$ is the $\mathbf{F}^{||}$ of the last step. This criterion is designed to prevent the dimer from going down hill to the IS/FS or from trapping around the inflection point and, thus, enable multiple geometry relaxation steps. The idea behind this can be described as follows: When the dimer is between an IS (or FS) and an inflection point where the curvature is positive, the dimer will translate toward the inflection point to maximize the parallel force. As the dimer passes through the inflection point, the parallel force drops, and the translation is stopped due to eq 13. After the inflection point, the recalculated curvature is negative, and the dimer will translate toward a TS to minimize the parallel force according to eq 13. In the case of systems with multiple inflection points along the reaction coordinate, the algorithm works similarly, as the parallel force has the local maximum or minimum at the inflection points. A translation step will be stopped similarly whenever an inflection point is passed. This will gradually lead the dimer away from the inflection points until the TS is located. To avoid the trapping at the shoulder point in a flat PES, it is essential to ensure the curvature of the located TS to be negative enough.

Furthermore, we found that the TS-searching trajectory (translation trajectory) could also be optimized by modifying the translational force. For illustration purposes, we demonstrated the TS-searching method on the simple PES mentioned above with the following program:

(i) Rotate dimer to identify the local curvature (in fact this can be done analytically in the 2D PES).

(ii) Perform the translational moves with the force according to eq 14, where a factor $\lambda$ (between 0 and 1) on $\mathbf{F}^{\parallel}$ is introduced. The choice of the value of $\lambda$ in real systems will be discussed in the next subsection. Obviously, when $\lambda$ equals to 1, eq 14 is the same as eq 3 for $C < 0$.

$$\mathbf{F}_{\mathrm{tran}} = \mathbf{F}^{\perp} - \lambda \mathbf{F}^{\parallel} \tag{14}$$

(iii) Terminate the translation if the condition of eq 13 is reached or if $|\mathbf{F}^{\perp}| < 10^{-5}$. For the $\lambda$=0 case, one additional move with the force being $-0.1\mathbf{F}^{\parallel}$ is carried out.

(iv) Repeat i−iii until $|\mathbf{F}| < 10^{-5}$.

(v) Check the final result by calculating the curvature of the converged state. If the curvature is close to zero, then it implies that the search may converge to a “shoulder state”. In such a case, we need to guess a better initial condition and to restart the search.

The effect of $\lambda$ on the TS-searching trajectory can be seen clearly by comparing the dotted color curves in Figure 1. The red curve represents the condition of $\lambda = 1$. The blue one represents a trajectory with $\lambda = 0$. Although both the red and the blue curves can finally converge to the TS, they appear to be two extremes from the optimum searching path, the black line in the figure. If we adjust the value of $\lambda$ in between (0 and 1) in eq 14, we can obtain trajectories between the blue and the red curves, such as the orange ($\lambda = 0.25$), the magenta ($\lambda = 0.1$), and the green ($\lambda = 0.05$). Obviously, the magenta curve is the optimum path among the five curves.

We would like to address further the meaning of $\lambda$ from two aspects. Mathematically, the implementation of $\lambda$ effectively projects out a fraction of the forces at the direction defined by the dimer $\hat{\mathbf{N}}$. By doing this, the minimization of the other degrees of freedom is of higher priority compared to that of the $\hat{\mathbf{N}}$ direction. As shown in the Figure 1 blue curve ($\lambda = 0$), the searching trajectory is toward and along MEP. From a practical point of view, the identified dimer $\hat{\mathbf{N}}$ direction is local, but the translation along $-\mathbf{F}^{\parallel}$ ($\lambda = 1$) is typically at a long step size (e.g., 0.2 Å). As a result, the translation with $\lambda = 1$ on a corrugated PES could be unstable, since the walk against the force may lead to the divergence. Therefore, the scaling of $\mathbf{F}^{\parallel}$ is desirable for safety.

*2.2.2. Implementation in Real Molecular Systems.* In a real system of $3N$ degrees of freedom, eq 14 can be applied similarly for the $C < 0$ regions. The value of $\lambda$ can be conveniently chosen as a set of values, as shown in eq 15, according to the rms value of $\mathbf{F}^{\parallel}$ ($|\mathbf{F}^{\parallel}|$) at the beginning of each translation ($\lambda$ is unchanged in one translation). It should be mentioned that by applying $\lambda$ values in between 0 and 1, the TS searching is actually constrained, and the algorithm becomes more stable because the trajectory is closer to MEP. This will be demonstrated in Section 3.

$$\mathbf{F}_{\mathrm{tran}} = \mathbf{F}^{\perp} - \lambda \mathbf{F}^{\parallel} \quad (C < 0); \quad \lambda = \begin{cases} 0.1 & |\mathbf{F}^{\parallel}| \in [2, +\infty) \\ 0.25 & |\mathbf{F}^{\parallel}| \in [1, 2) \\ 0.5 & |\mathbf{F}^{\parallel}| \in [0.5, 1) \\ 1.0 & |\mathbf{F}^{\parallel}| \in [0, 0.5) \end{cases} \quad (\mathrm{eV/Å}) \tag{15}$$

$$\mathbf{F}_{\mathrm{tran}} = \begin{cases} 0.5{\cdot}\mathbf{F}^{\perp} - \mathbf{F}^{\parallel} & (|\mathbf{F}^{\perp}|<2) \\ \mathbf{F}^{\perp} - 0.5{\cdot}\mathbf{F}^{\parallel} & (|\mathbf{F}^{\perp}|>2) \end{cases} (\mathrm{eV/Å}) \quad (C > 0) \tag{16}$$

$$\begin{cases} |\mathbf{F}_i^{\parallel}| > |\mathbf{F}_{i-1}^{\parallel}| & (C < 0) \\ |\mathbf{F}_i^{\parallel}| < |\mathbf{F}_{i-1}^{\parallel}| \quad \text{or} \quad |\mathbf{F}_i^{\perp}| > |\mathbf{F}_{i-1}^{\perp}| & (C > 0) \end{cases} \tag{17}$$

For a similar reason, we also introduce a prefactor 0.5 to the $\mathbf{F}^{\perp}$ or $\mathbf{F}^{\parallel}$ to optimize the searching trajectory in the $C >$ 0 regions, as expressed in eq 16. Equation 16 is designed to reduce the vertical force, preferentially when the force is too large (e.g., $|\mathbf{F}^{\perp}| > 2$ eV/Å), which drags the image toward MEP and to maximize the parallel force when the force is small enough ($|\mathbf{F}^{\perp}| < 2$), which drags the image toward the TS. In our work, the vertical force is always applied in the positive-curvature region (c.f., eq 3) to relax the structure to MEP. Accordingly, a criterion as eq 17 is utilized for the termination of translational move. Equation 17 ensures the minimization of the vertical force together with the maximization of the parallel force during the dimer translation in the $C > 0$ regions.

*2.2.3. Performance of Boyden Algorithm at the Negative-Curvature Region.* While the Broyden technique proved to be applicable for minimization problems when the actual Hessian is positive definite, we here utilized the Broyden technique to locate TS, where the actual or the real Hessian matrix has one negative eigenvalue. To enable the Broyden technique to locate the TS as required, the key is to reverse $\mathbf{F}^{\parallel}$ as $-\mathbf{F}^{\parallel}$, which effectively changes the curvature at this degree of freedom to be positive. According to eq 1, where $C = -\mathrm{d}\mathbf{F}^{\parallel}/\mathrm{d}x$, it is clear that the sign of the $C$ associated with the parallel force is changed by inverting $\mathbf{F}^{\parallel}$. Such transformed forces fed into Broyden enables the update of the Jacobian matrix, according to the modified

***Table 1.*** Illustration of the Modified Broyden Algorithm in Finding the Minimum Point (IS) and the Saddle Point (TS) on a Simple Five Dimension PES System[a]

| iteration | $\|\bar{\mathbf{x}}\|$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|---|---|---|---|---|---|---|
| IS | | | | | | |
| 0 | 2.8591 | −0.1000 | −0.2000 | −0.3000 | −0.4000 | −0.5000 |
| 1 | 0.7604 | −1.0950 | −1.1801 | −1.2553 | −1.3211 | −1.3776 |
| 2 | 0.3432 | −1.7772 | −1.7473 | −1.7174 | −1.6891 | −1.6635 |
| 3 | 0.0134 | −1.5612 | −1.5641 | −1.5660 | −1.5674 | −1.5683 |
| 4 | 0.0041 | −1.5742 | −1.5727 | −1.5718 | −1.5714 | −1.5711 |
| 5 | 0.0000 | −1.5708 | −1.5708 | −1.5708 | −1.5708 | −1.5708 |
| TS | | | | | | |
| 0 | 2.8591 | −0.1000 | −0.2000 | −0.3000 | −0.4000 | 0.5000 |
| 1 | 0.7604 | −1.0950 | −1.1801 | −1.2553 | −1.3211 | 1.3776 |
| 2 | 0.3432 | −1.7772 | −1.7473 | −1.7174 | −1.6891 | 1.6635 |
| 3 | 0.0134 | −1.5612 | −1.5641 | −1.5660 | −1.5674 | 1.5683 |
| 4 | 0.0041 | −1.5742 | −1.5727 | −1.5718 | −1.5714 | 1.5711 |
| 5 | 0.0000 | −1.5708 | −1.5708 | −1.5708 | −1.5708 | 1.5708 |

[a] The $|\bar{\mathbf{x}}|$ measures the difference of the current position with respect to the converged position.

Broyden algorithm, to be positive definite. In our implementation of the multiple translation steps at $C < 0$ region, once the Jacobian matrix becomes nonpositive definite, which is an indication that the reversed $\mathbf{F}^{\|}$ does not fully quench the negative mode, the translation should always be terminated, and a new iteration to rotate the dimer will start. We would like to emphasize that both CG and L-BFGS methods have been utilized in the previous versions of the dimer method,[18–21] which demonstrated that it is, in principle, possible to utilize the minimization techniques to locate the TS with the reversed parallel force.

To illustrate this more clearly, we show the searching for the minimum point and the saddle point with the modified Broyden algorithm on a simple five dimension PES $E = \sum_{i=1}^{5} E_i$, where $E_i = \sin(x_i)$. From Table 1, we can see that when the initial $x_i$ are given as (−0.1,−0.2,−0.3,−0.4,and −0.5), the Broyden can locate the minimum ($-\pi/2$, $-\pi/2$, $-\pi/2$, $-\pi/2$, and $-\pi/2$), as expected in five steps where $E = -5$. Next, we changed the initial $x_i$ to (−0.1,−0.2,−0.3, −0.4, and 0.5) where the last dimension is close to the maximum (the Hessian has one negative eigenvalue). By reversing the force at this dimension only ($-F_5$) but keeping the right force at the other dimensions, the Broyden can locate both the maximum in this dimension and the minima in the other dimensions, ($-\pi/2$, $-\pi/2$, $-\pi/2$, $-\pi/2$, and $\pi/2$). The efficiency is the same as the location of the minima, and the Jacobian matrix from Broyden has been verified to be positive definite.

## 3. Results and Discussion

To test the efficiency of our approach, we first chose the Baker reaction system[43] as the testing examples, which contains 25 different chemical reactions as listed in Table 2. The same system has been utilized to test the modified dimer method, according to the reference.[20,21] In this work, four different algorithms, denoted as (**00**), (**01**), (**10**), and (**11**), were implemented to test the individual efficiency of the rotation and translation parts. The algorithms (**00**): the CG dimer method, as reported in refs 20 and 37; (**01**): the same as (**00**) except that the translation uses our Broyden translation algorithm; (**10**): the same as (**00**) except that the rotation uses the constrained Broyden rotation algorithm; and (**11**): our new method.

For all the reactions studied, we started from the same guess structure ($\mathbf{R}_{GS}$), as suggested by Baker.[43] The initial mode for the dimer was set as $\hat{\mathbf{N}}_{ini} = \mathbf{R}_{GS} - \mathbf{R}_{IS}$, except for reaction 10 (*s*-tetrazine → 2HCN + N$_2$), where $\hat{\mathbf{N}}_{ini} = \mathbf{R}_{FS} - \mathbf{R}_{IS}$ was used. All calculations were performed using the SIESTA package[44] with numerical double-ζ-polarization basis set[45,46] at the density functional theory level, in which the GGA-PBE exchange−correlation functional was utilized.[47] The rotation stops if the $|\Delta\mathbf{F}^{\perp}|$ is lower than 0.1 eV/Å. The translation is terminated by the condition of eq 17. The TS searching is converged if the maximum force on each freedom (max $|\mathbf{F}|$) is below 0.1 eV/Å. All the determined TSs have been checked with the literature structure[43] to ensure that the correct TS is identified. In Table 2, our results on the total calculation steps (the number of energy and gradient calculations) are shown. The "###" sign in the table represents that the corresponding method either produces the wrong TS or diverges after 400 steps.

From Table 2, we found that the average number of steps by using (**00**) method is 91.7 for the 21 reactions with located correct TS, and the method fails in three reactions 10, 11, and 15. In the (**11**) method, the average number of steps is reduced to 35.3, which is about 40% of the (**00**) method (the results of reactions 10, 11, and 15 are not included for comparison between methods here after). Importantly, while the (**00**) method fails in three reactions, all the desired TSs have been located using the (**11**) method. By comparing the four methods, we can see that our algorithms on both the rotation and the translation can help to increase the efficiency. Specifically, the modification to the rotation and the translation only can reduce the step numbers by 23 and 51%, respectively. These will be elaborated further in the following.

**3.1. Efficiency of Rotation.** Table 3 compares the efficiency of the rotation between the (**00**) and (**10**) methods. We can see that the total iteration numbers of the two methods are nearly identical (one iteration contains all energy and gradient calculation steps between two consecutive rotations, including those for both rotation and translation). However, the average energy and gradient calculation steps in each rotation are reduced in the (**10**) method, where the Broyden method is utilized for the dimer rotation. On average, the (**00**) method needs 5.0 energy and gradient calculation steps per rotation, and the (**10**) method needs 3.5 energy and gradient calculations per rotation, which means that the Broyden rotation algorithm can save about 30% computational load. Compared with the extrapolation method proposed by Kästner and Sherwood,[21] that can save about 11% computational load in rotation, the constrained Broyden rotation shows the better performance.

One additional feature of Broyden rotation is that it tends to find a local normal mode by minimizing the rotational force and, thus, is more sensitive to the initial mode provided as the input, which is most naturally determined from the guess initial structure as $\hat{\mathbf{N}}_{ini} = \mathbf{R}_{GS} - \mathbf{R}_{IS}$, or from the final state as $\hat{\mathbf{N}}_{ini} = \mathbf{R}_{FS} - \mathbf{R}_{IS}$ in complex PES reactions. This appears to be advantageous in Baker reactions. By utilizing

***Table 2.*** The Energy and Gradient Calculation Steps from Four Different Methods in the TS Location of Baker Reactions[a]

| | | (**00**) | (**10**) | | (**01**) | | (**11**) | |
|---|---|---|---|---|---|---|---|---|
| | system | steps | steps | radio | steps | radio | steps | radio |
| 1 | HCN → HNC | 71 | 60 | 85% | 66 | 93% | 32 | 45% |
| 2 | HCCH → $CCH_2$ | 93 | 67 | 72% | 52 | 56% | 40 | 43% |
| 3 | $H_2CO$ → $H_2$+CO | 62 | 46 | 74% | 39 | 63% | 31 | 50% |
| 4 | $CH_3O$ → $CH_2OH$ | 68 | 51 | 75% | 37 | 54% | 29 | 43% |
| 5 | ring-opening cyclopropyl | 88 | 73 | 83% | 41 | 47% | 32 | 36% |
| 6 | bicyclo110 butane TS1 | 96 | 80 | 83% | 36 | 38% | 40 | 42% |
| 7 | bicyclo110 butane TS2 | 144 | 101 | 70% | 56 | 39% | 46 | 32% |
| 8 | $\beta$-(formyloxy) ethyl | 88 | 67 | 76% | 24 | 27% | 19 | 22% |
| 9 | parent Diels−Alder | 132 | 122 | 92% | 73 | 55% | 52 | 39% |
| 10 | *s*-tetrazine → 2HCN + $N_2$ | ### | 101 | − | ### | − | 59 | − |
| 11 | rotational TS in butadiene | ### | 119 | − | 51 | − | 29 | − |
| 12 | $H_3CCH_3$ → $H_2CCH_2$ + $H_2$ | 77 | 63 | 82% | 41 | 53% | 32 | 42% |
| 13 | $H_3CCH_2F$ → $H_2CCH_2$ + HF | 89 | 64 | 72% | 49 | 55% | 41 | 46% |
| 14 | $H_2CCHOH$ → $H_3CCHO$ | 141 | 116 | 82% | 104 | 74% | 101 | 72% |
| 15 | HCOCl → HCl + CO | ### | 96 | − | 151 | − | 109 | − |
| 16 | $H_2O$ + $PO_3^-$ → $H_2PO_4$ | 181 | 84 | 46% | 45 | 25% | 35 | 19% |
| 17 | Claisen rearrangement | 127 | 92 | 72% | 44 | 35% | 37 | 29% |
| 18 | silylene insertion | 70 | 58 | 83% | 29 | 41% | 26 | 37% |
| 19 | HNCCS → HNC + CS | 47 | 36 | 77% | 28 | 60% | 25 | 53% |
| 20 | $HCONH_3^+$ → $NH_4^+$ + CO | 113 | 80 | 71% | 46 | 41% | 35 | 31% |
| 21 | rotational TS in acrolein | 67 | 58 | 87% | 22 | 33% | 17 | 25% |
| 22 | HCONHOH → HCOHNHO | 63 | 46 | 73% | 32 | 51% | 25 | 40% |
| 23 | HNC + $H_2$ → $H_2CNH$ | 59 | 50 | 85% | 30 | 51% | 25 | 42% |
| 24 | $H_2CNH$ → $HCNH_2$ | 109 | 86 | 79% | 46 | 42% | 40 | 37% |
| 25 | $HCNH_2$ → HCN + $H_2$ | 32 | 27 | 84% | 16 | 50% | 17 | 53% |
| | average | 91.7 | 69.4 | 77% | 43.5 | 49% | 35.3 | 40% |

[a] The ratio is the step number with the method referred divided by that with (**00**) method. The average values listed do not count the reactions 10, 11, and 15, where the (**00**) or (**01**) method fails to locate the TS.

***Table 3.*** Comparison Between (**00**) and (**10**) Methods for TS Location of Baker Reactions (Numbered from 1 to 25, as in Table 2)[a]

| | method | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | av |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| iter | (**00**) | 9 | 11 | 8 | 9 | 14 | 18 | 24 | 12 | 19 | ### | ### | 12 | 13 | 20 | ### | 12 | 16 | 13 | 7 | 18 | 13 | 9 | 8 | 17 | 5 | 13.0 |
| | (**10**) | 8 | 12 | 8 | 9 | 14 | 17 | 24 | 12 | 22 | 21 | 27 | 12 | 11 | 19 | 15 | 13 | 16 | 13 | 7 | 14 | 15 | 8 | 8 | 17 | 5 | 12.9 |
| av. rot | (**00**) | 6 | 6 | 6 | 6 | 4 | 3 | 4 | 5 | 5 | ### | ### | 4 | 5 | 5 | ### | 13 | 6 | 3 | 4 | 5 | 3 | 5 | 5 | 4 | 4 | 5.0 |
| | (**10**) | 6 | 4 | 4 | 3 | 3 | 3 | 2 | 4 | 4 | 5 | 2 | 3 | 4 | 4 | 4 | 4 | 4 | 2 | 3 | 4 | 2 | 4 | 4 | 3 | 3 | 3.5 |

[a] Listed are the average rotation steps (av. rot) per iteration and the total iteration number (iter). The average values (av) listed do not count the reactions 10, 11, and 15, where the (**00**) method fails to locate TS.

***Table 4.*** Comparison between (**00**) and (**01**) Method for TS location of Baker Reactions (Numbered from 1 to 25, as in Table 2)[a]

| | method | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | av |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| iter | (**00**) | 9 | 11 | 8 | 9 | 14 | 18 | 24 | 12 | 19 | ### | ### | 12 | 13 | 20 | ### | 12 | 16 | 13 | 7 | 18 | 13 | 9 | 8 | 17 | 5 | 13.0 |
| | (**01**) | 6 | 4 | 3 | 3 | 2 | 2 | 3 | 1 | 3 | ### | 4 | 3 | 3 | 6 | 11 | 3 | 2 | 1 | 1 | 2 | 1 | 2 | 2 | 3 | 1 | 2.6 |
| trans | (**00**) | 18 | 22 | 16 | 16 | 28 | 36 | 48 | 24 | 38 | ### | ### | 24 | 26 | 40 | ### | 24 | 32 | 26 | 14 | 36 | 26 | 18 | 16 | 34 | 10 | 26.0 |
| | (**01**) | 18 | 18 | 13 | 18 | 22 | 18 | 25 | 12 | 31 | ### | 31 | 17 | 21 | 62 | 43 | 15 | 35 | 13 | 12 | 20 | 8 | 12 | 10 | 14 | 6 | 19.1 |
| rot | (**00**) | 55 | 69 | 46 | 53 | 60 | 62 | 96 | 64 | 88 | ### | ### | 53 | 67 | 103 | ### | 155 | 95 | 44 | 27 | 85 | 37 | 47 | 43 | 75 | 22 | 65.7 |
| | (**01**) | 48 | 34 | 26 | 19 | 19 | 18 | 31 | 12 | 42 | ### | 20 | 24 | 28 | 42 | 108 | 30 | 9 | 16 | 16 | 26 | 14 | 20 | 20 | 32 | 10 | 24.4 |

[a] Listed are the total iteration number (iter) and the total number of energy and gradient calculations in translation (trans) and in rotation (rot). The average values (av) listed do not count the reactions 10, 11, and 15, where the (**00**) or (**01**) method fails to locate the TS.

the constrained Broyden rotation algorithm, the (**10**) method, we show that all the 25 reactions can converge to the correct TS, while the (**00**) method fails in three reactions.

**3.2. Efficiency of Translation.** Table 4 compares the efficiency of the translation between the (**00**) and (**01**) methods. We see that our translation algorithm can decrease the total iteration number greatly, where the iteration number in (**01**) is only 20% of that in (**00**). This is largely because the current translation algorithm carries out multiple translational moves along a better trajectory and can move much longer in distance per iteration compared to that of the (**00**) method. The higher efficiency in translation, in turn, reduces effectively the number of rotation needed. As the rotation steps involve heavily the energy and gradient calculations, the reduction in the rotation number can save about 63% computational load, as seen from Table 4. Besides, the translation part can also save 27% computational load, which can be attributed to the application of constrained Broyden algorithm.

In order to see more clearly how the multiple translation works, we have generated a three-dimensional (3D) diagram to trace the searching trajectory, as shown in Figure 2, where the *x*-axis is along IS−FS vector by setting IS at (0,0,0) and FS at (1,0,0), the *y*-axis is determined by the TS at ($x_{TS}$,1,0)

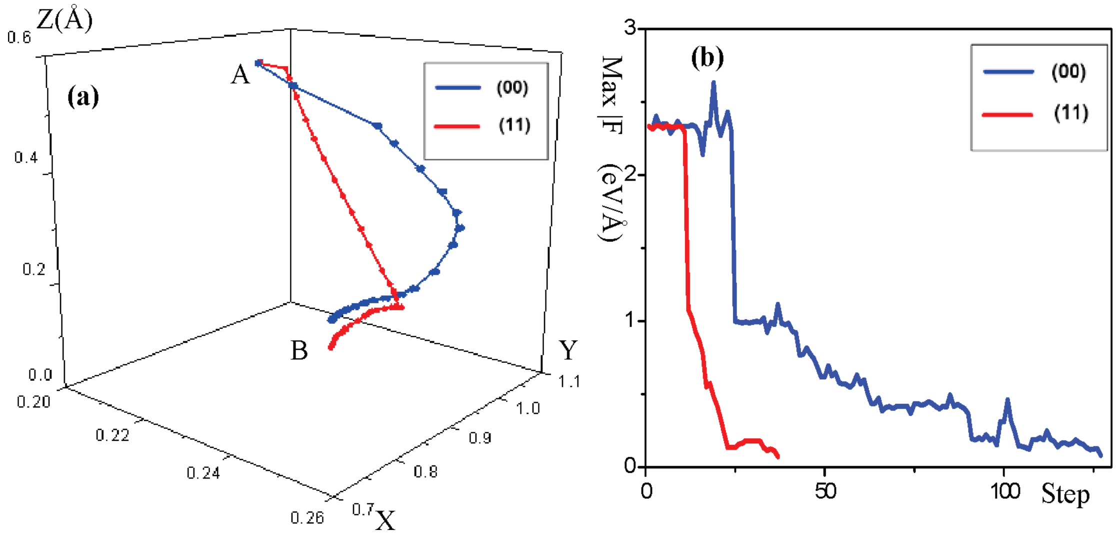


*Figure 2.* (a): 3D trajectory of TS searching in Claisen rearrangement reaction. The *x*-axis is along IS−FS vector by setting IS at (0,0,0) and FS at (1,0,0); the *y*-axis is determined by the TS at ($x_{TS}$,1,0) where $x_{TS}$ is the projection of TS on the IS−FS vector; and the *z*-axis is the direction perpendicular to the IS−TS−FS plane. The meaning of the labels are as follows: A is the initial point (guess structure), and B is the end point (located TS). (b): The plot showing the maximum force on each freedom (max |**F**|) in the system during the TS searching.

where $x_{TS}$ is the projection of TS on the IS−FS vector, and the *z*-axis is the direction perpendicular to the IS−TS−FS plane. A large *z* value of a structure would imply that the structure is far away from the reaction plane (e.g., high in energy) and, thus, is not desirable in TS searching.

Using Claisen rearrangement ($CH_2CHCH_2CH_2CHO$ → $CH_2CHOCH_2CHCH_2$) (Figure 2a and b) reaction as the example, we can see two distinct trajectories by using (**00**) and (**11**), as shown by the blue and the red curve, respectively. In the 3D figure, the *x*-axis is along IS−FS vector by setting IS at (0,0,0) and FS at (1,0,0); the *y*-axis is determined by the TS at ($x_{TS}$,1,0), where $x_{TS}$ is the projection of TS on the IS−FS vector; and the *z*-axis is the direction perpendicular to the IS−TS−FS plane. The unit vector $\hat{\mathbf{X}}$ and $\hat{\mathbf{Y}}$and the *z* value of an image $\mathbf{R}(z_R)$ are defined in eqs 18−20.

$$\hat{\mathbf{X}} = \frac{\mathbf{R}^{FS} - \mathbf{R}^{IS}}{|\mathbf{R}^{FS} - \mathbf{R}^{IS}|} \tag{18}$$

$$\hat{\mathbf{Y}} = \frac{(\mathbf{R}^{TS} - \mathbf{R}^{IS}) - [(\mathbf{R}^{TS} - \mathbf{R}^{IS})\bullet\hat{\mathbf{X}}]\cdot\hat{\mathbf{X}}}{|(\mathbf{R}^{TS} - \mathbf{R}^{IS}) - [(\mathbf{R}^{TS} - \mathbf{R}^{IS})\bullet\hat{\mathbf{X}}]\cdot\hat{\mathbf{X}}|} \tag{19}$$

$$z_R = |(\mathbf{R} - \mathbf{R}^{IS}) - [(\mathbf{R} - \mathbf{R}^{IS})\bullet\hat{\mathbf{X}}]\cdot\hat{\mathbf{X}} - [(\mathbf{R} - \mathbf{R}^{IS})\bullet\hat{\mathbf{Y}}]\cdot\hat{\mathbf{Y}}| \tag{20}$$

Peters et al.[48] and Branduardi et al.[49] have suggested methods to measure the distance of a structural image along the reaction coordinate and the displacement from the MEP. These methods need the information of the whole MEP, which is, however, not known practically from the dimer approach. The 3D plot in Figure 2a is, thus, utilized for the comparison of the length of different searching trajectories. Figure 2a shows that there is a turning point in the red curve where the translation of the first iteration meets the termination criterion and where the rotation in the second iteration starts. It appears that the first rotation guides the dimer toward the reaction plane, and the second rotation leads to the exact TS. The red trajectory is shorter than the blue one. The length of the trajectory is 0.99 and 1.40 Å for the red and blue curves, respectively. The red trajectory is close to the straight line distance between A and B ($|\mathbf{R}_A - \mathbf{R}_B|$), 0.70 Å. With only two rotations and 37 steps, the (**11**) method identifies the TS, whereas there are 16 rotations and 127 steps with the (**00**) method, as compared clearly in Figure 2b, where the maximum force on each freedom in the system is traced during the TS searching. It shows that an efficient translational move can indeed be realized with the approximate mode. The termination criterion for the dimer translation is the key to achieve high efficiency for TS location.

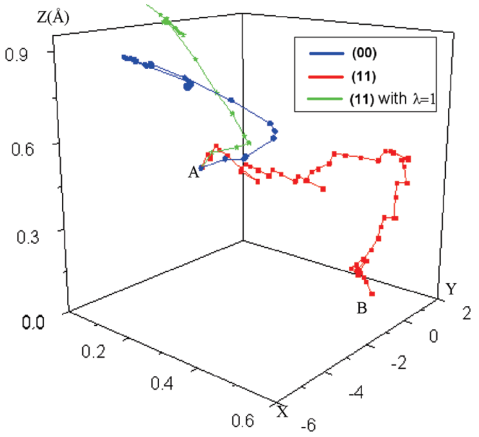


*Figure 3.* 3D trajectory of TS searching in HCOCl → HCl + CO reaction. The meaning of axis and labels are the same as those in Figure 2a.

Finally, we would like to address the effect of $\lambda$ in $\mathbf{F}_{tran}$ on the TS-searching trajectory. Using the reaction HCOCl → HCl + CO (Figure 3) as the example, we have plotted three trajectories: the blue curve with (**00**) method, the red curve with (**11**) method, and the green curve with (**11**) method but with $\lambda = 1$. Only the red curve finds the TS. In the blue and green curves ($\lambda$ being 1, i.e., by acting $-\mathbf{F}^{||}$

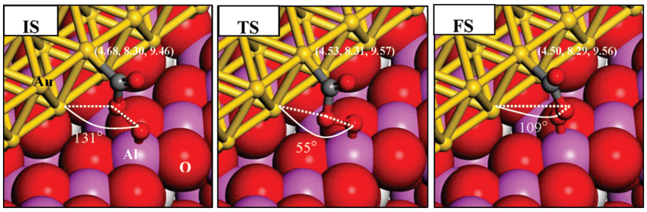


***Figure 4.*** Reaction snapshots for a OCOO rotation at the Au/$\gamma$-$Al_2O_3$ interface (O, red; Al, purple; H, white; Au, yellow; and C, gray). The Cartesian coordinates of the Au atom that bonds with C is indicated in parentheses.

directly), the structure goes quickly away from the reaction plane due to the too large stretching against the force at the parallel dimer $\hat{\mathbf{N}}$ direction. The trajectory can not be directed back toward the reaction plane because the rotation of the dimer later is localized to the wrong mode. Following such a wrong mode, the TS searching will diverge or lead to undesired TS. Therefore, it can be concluded that with an optimized $\lambda$ factor to $\mathbf{F}^{\|}$, the structure during optimization may leave the unfavorable high-energy region rapidly, and the algorithm is, thus, more stable and efficient.

**3.3. Application in a Large Heterogeneous Catalytic System.** For its no need of the Hessian, the dimer method is well suited for complex reactions occurring in a metallic system, which is not ideal to treat with Gaussian basis sets. Here, we further illustrate our methods in finding a TS involved in the CO oxidation on a Au/$\gamma$-$Al_2O_3$ model catalyst, where a two-layer Au strip is deposited on the (100) surface of $\gamma$-$Al_2O_3$. The CO oxidation on Au supported on oxides has been studied recently,[50,51] where a bimolecular pathway CO + $O_2$→OCOO occurring at the Au and the oxide interface was identified. The OCOO can further decompose into $CO_2$ and adsorbed O. For the OCOO intermediate at the Au/oxide interface, it has two isomeric structures which are connected by the rotation of OO that is beneath CO. A TS of the rotation can be identified, as shown in the Figure 4, together with the IS and the FS. In the system, 90 degrees of freedom are relaxed, including those of the Au and the top layer $\gamma$-$Al_2O_3$ atoms. It should be mentioned that the substrate during the reaction is not rigid with a large displacement from IS to TS, as indicated by the coordinate of the Au atom labeled in Figure 4. The rotation TS is just the kind of TS that is difficult to locate using the CBM method by fixing a certain bond distance.

The above four methods have been applied to locate the TS and the required energy and gradient calculations steps are 511, 353, 197, and 95 for the (**00**), (**10**), (**01**) and (**11**) methods, respectively. We see that, in such a large system, the (**11**) method can also achieve the highest performance, where about 80% CPU time is saved compared to that of the original (**00**) method. The result demonstrates that the dimer rotation and the translation by utilizing the Broyden approach scales well to the large system with many degrees of freedom. For such a large system, the computation of exact Hessian becomes extremely demanding if a numerical algorithm via finite difference method is utilized.

It is noticed that even with the current improvement on efficiency, the average step number of TS location in the Baker system, in this work, is still two times more than that reported in the original Baker paper, where the P-RFO method (Powell scheme for Hessian update based on an initial analytic Hessian) is utilized. Nevertheless, we would like to emphasize that the dimer method could be the better choice especially when the system is large and when the second derivatives are not available cheaply. This has been addressed previously, as demonstrated by Heyden et al.[20] and Kästner et al.,[21] the original dimer method is already preferable to P-RFO when the Hessian is not cheaply available. We, therefore, believe that the method reported here could be the best choice for locating TSs of large complex reaction systems. Finally, it should be mentioned that compared with the chain-of-states methods, the surface-walking methods, including the dimer method, generally have the so-called "dead-end valley" problem to miss TS off. The dimer method, therefore, also requires a reasonably guessed initial structure in order to locate successfully the desired TS in complex reaction systems. To overcome the problem, Peters et al. has developed a method, which can "teach" saddle search algorithms to locate multiple reaction pathways.[48]

## 4. Conclusion

This work combines the constrained Broyden minimization method with the dimer method for locating TS without the need of Hessian. New algorithms are designed with the aim to maximally cut the rotation steps and to increase the length of translational move. Our method was implemented and tested in the Baker reaction system and also in a large heterogeneous catalytic system, which shows the enhanced stability and the higher efficiency. We demonstrate that the atomic force parallel to the dimer direction can be damped to increase the stability of the TS searching and to shorten the searching trajectory.

**Acknowledgment.** This work is supported by the National Science Foundation (NSF) of China (20825311, 20773026, and 20721063), the Science and Technology Community Shanghai Municipality (08DZ2270500), and the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institute of Higher Learning.

### References

(1) Henkelman, G.; Jonsson, H. *J. Chem. Phys.* **2000**, *113*, 9978.

(2) Henkelman, G.; Uberuaga, B. P.; Jonsson, H. *J. Chem. Phys.* **2000**, *113*, 9901.

(3) Mills, G.; Jonsson, H. *Phys. Rev. Lett.* **1994**, *72*, 1124.

(4) Sheppard, D.; Terrell, R.; Henkelman, G. *J. Chem. Phys.* **2008**, *128*, 134106.

(5) Elber, R.; Karplus, M. *Chem. Phys. Lett.* **1987**, *139*, 375.

(6) Trygubenko, S. A.; Wales, D. J. *J. Chem. Phys.* **2004**, *120*, 7820.

(7) Trygubenko, S. A.; Wales, D. J. *J. Chem. Phys.* **2004**, *120*, 2082.

(8) Koslover, E. F.; Wales, D. J. *J. Chem. Phys.* **2007**, *127*, 134102.

(9) Carr, J. M.; Trygubenko, S. A.; Wales, D. J. *J. Chem. Phys.* **2005**, *122*, 234903.

(10) Peters, B.; Heyden, A.; Bell, A. T.; Chakraborty, A. *J. Chem. Phys.* **2004**, *120*, 7877.

(11) E, W. N.; Ren, W. Q.; Vanden-Eijnden, E. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2002**, *66*, 052301.

(12) Simons, J.; Jorgensen, P.; Taylor, H.; Ozment, J. *J. Phys. Chem.* **1983**, *87*, 2745.

(13) Khait, Y. G.; Puzanov, Y. V. *J. Mol. Struct. (THEOCHEM)* **1997**, *398*, 101.

(14) Cerjan, C. J.; Miller, W. H. *J. Chem. Phys.* **1981**, *75*, 2800.

(15) Baker, J. *J. Comput. Chem.* **1986**, *7*, 385.

(16) Munro, L. J.; Wales, D. J. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1999**, *59*, 3969.

(17) Kumeda, Y.; Wales, D. J.; Munro, L. J. *Chem. Phys. Lett.* **2001**, *341*, 185.

(18) Henkelman, G.; Jonsson, H. *J. Chem. Phys.* **1999**, *111*, 7010.

(19) Olsen, R. A.; Kroes, G. J.; Henkelman, G.; Arnaldsson, A.; Jonsson, H. *J. Chem. Phys.* **2004**, *121*, 9776.

(20) Heyden, A.; Bell, A. T.; Keil, F. J. *J. Chem. Phys.* **2005**, *123*, 224101.

(21) Kaestner, J.; Sherwood, P. *J. Chem. Phys.* **2008**, *128*, 014106.

(22) Poppinger, D. *Chem. Phys. Lett.* **1975**, *35*, 550.

(23) Wang, H. F.; Liu, Z. P. *J. Am. Chem. Soc.* **2008**, *130*, 10996.

(24) Powell, M. J. D. *Mathematical Programming* **1971**, *1*, 26.

(25) Schlegel, H. B. *J. Comput. Chem.* **1982**, *3*, 214.

(26) Culot, P.; Dive, G.; Nguyen, V. H.; Ghuysen, J. M. *Theor. Chim. Acta* **1992**, *82*, 189.

(27) Fletcher, R. *Practical Methods of Optimization*, 2nd ed.; Wiley: Chichester, U.K., 1982; Vol. 1.

(28) Dennis, J. E.; Schnabel, R. B. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*; Prentice-Hall: Englewood Cliffs, NJ, 1983.

(29) Bofill, J. M. *J. Comput. Chem.* **1994**, *15*, 1.

(30) Anglada, J. M.; Besalu, E.; Bofill, J. M.; Rubio, J. *J. Math. Chem.* **1999**, *25*, 85.

(31) Bofill, J. M.; Comajuan, M. *J. Comput. Chem.* **1995**, *16*, 1326.

(32) Bofill, J. M. *Chem. Phys. Lett.* **1996**, *260*, 359.

(33) Anglada, J. M.; Bofill, J. M. *J. Comput. Chem.* **1998**, *19*, 349.

(34) Bofill, J. M.; Anglada, J. M. *Theor. Chem. Acc.* **2001**, *105*, 463.

(35) Besalu, E.; Bofill, J. M. *Theor. Chem. Acc.* **1998**, *100*, 265.

(36) Schlegel, H. B. *J. Comput. Chem.* **2003**, *24*, 1514.

(37) *The FORTRAN code for the cg-dimer method*; Henkelman Research Group: Austin, TX; http://theory.cm.utexas.edu/henkelman. Accessed February 21, 2010.

(38) Kresse, G.; Furthmuller, J. *Comput. Mater. Sci.* **1996**, *6*, 15.

(39) Broyden, C. G. *Mathematics of Computation* **1965**, *19*, 557.

(40) Vanderbilt, D.; Louie, S. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1984**, *30*, 6118.

(41) Johnson, D. D. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *38*, 12807.

(42) Fischer, T. H.; Almlof, J. *J. Phys. Chem.* **1992**, *96*, 9768.

(43) Baker, J.; Chan, F. R. *J. Comput. Chem.* **1996**, *17*, 888.

(44) Soler, J. M.; Artacho, E.; Gale, J. D.; Garcia, A.; Junquera, J.; Ordejon, P.; Sanchez-Portal, D. *J. Phys.: Condens. Matter* **2002**, *14*, 2745.

(45) Junquera, J.; Paz, O.; Sanchez-Portal, D.; Artacho, E. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2001**, *64*, 235111.

(46) Anglada, E.; Soler, J. M.; Junquera, J.; Artacho, E. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2002**, *66*, 205101.

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

(48) Peters, B.; Liang, W. Z.; Bell, A. T.; Chakraborty, A. *J. Chem. Phys.* **2003**, *118*, 9533.

(49) Branduardi, D.; Gervasio, F. L.; Parrinello, M. *J. Chem. Phys.* **2007**, *126*, 054103.

(50) Liu, Z. P.; Gong, X. Q.; Kohanoff, J.; Sanchez, C.; Hu, P. *Phys. Rev. Lett.* **2003**, *91*, 266102.

(51) Wang, C. M.; Fan, K. N.; Liu, Z. P. *J. Am. Chem. Soc.* **2007**, *129*, 2642.

CT9005147

Journal of Chemical Theory and Computation

# Metadynamics Simulations of Enantioselective Acylation Give Insights into the Catalytic Mechanism of *Burkholderia cepacia* Lipase

Luca Bellucci,*,† Teodoro Laino,‡,§ Andrea Tafi,*,† and Maurizio Botta†

*Dipartimento Farmaco Chimico Tecnologico, Università degli Studi di Siena, Via Aldo Moro 2, I-53100 Siena, Italy, Physikalisch Chemisches Institut, Universität Zürich, Winterthurerstrasse 190, CH-8057 Zürich Switzerland, and IBM Zurich Research Laboratory, Säumerstrasse 4, CH-8803 Rüschlikon Switzerland*

Received November 30, 2009

**Abstract:** The catalytic mechanism of *Burkholderia cepacia* lipase (BCL), which catalyzes the enantioselective hydrolysis of racemic esters of primary alcohols, was investigated by modeling the first stage of the enzymatic hydrolysis of (S/R)-2-methyl-3-phenyl-propanol (MPP) acetate, using molecular dynamics simulations in a mixed quantum mechanical/molecular mechanical (QM/MM) framework. The free energy surface of the enzyme acylation reaction was computed for both enantiomers. The simulations predict the existence of different reaction free energies that favor the (S)-enantiomer over the (R)-enantiomer by 5 kcal/mol. Analysis of the structural and dynamical aspects of the simulated reactions reveals an unforeseen reorganization of the catalytic triad in the (R)-MPP ester, driven by steric hindrance and involving the residues Asp264 and Glu289. Exploiting the different catalytic role of the above-mentioned acidic residues, we suggest a way to regulate the enantioselectivity of BCL by means of a few judicious point mutations that prevent the formation of the second catalytic triad used in the reaction with the (R)-enantiomer.

## 1. Introduction

Serine hydrolases are one of the largest and most diverse families of enzymes in higher eukaryotes. They comprise approximately 1% of the genes in the human genome, and because of their extensive usage in organic synthesis, they are the most investigated enzymes in pharmaceutical research.

Well-known members of this enzyme family include serine proteases such as α-chymotrypsin,[1,2] one of the first proteases to be revealed using X-ray crystallography,[3] esterases such as the acetylcholinesterase enzyme[4](AChE), which plays an important role in Alzheimer's disease,[5,6] and last but not least, lipases,[7] which are widely used for biotechnological applications.[8–12]

Although the natural function of lipases is to catalyze the hydrolysis of triacylglycerols, they also show high catalytic activity and unusual enantioselectivity toward a wide range of unnatural substrates.[13,14] These enzymes are widely used to separate racemic mixtures of chiral esters through hydrolysis or transesterification reactions, so that enantiomeric discrimination by lipases represents one of the most efficient biocatalytic strategies for producing enantiomerically pure pharmaceutical building blocks.[15–18]

Lipases share a characteristic catalytic mechanism with the remainder of serine hydrolases, involving the so-called catalytic triad consisting of the amino acids serine, histidine, and aspartic (or glutamic) acid. In addition to the catalytic triad, another important component of the active center of lipases is the oxyanion hole, a structural feature composed of hydrogen bond donors in the vicinity of the catalytic serine.[1,7,19,20]

Lipases work through a general mechanism peculiar to serine proteases,[1,7,20] known as the bi-bi ping-pong mecha-

* Corresponding author e-mail: bellucci14@unisi.it (L.B.); tafi@unisi.it (A.T.).

† Università degli Studi di Siena.

‡ Universität Zürich.

§ IBM Zurich Research Laboratory.

10.1021/ct900636w
Published on Web 03/11/2010

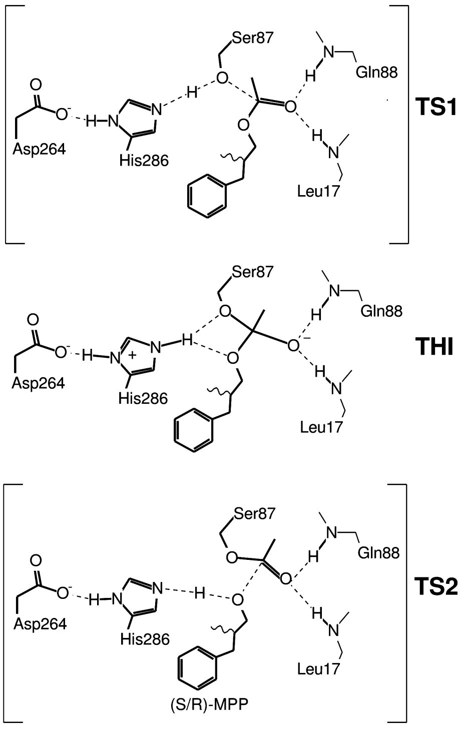


***Figure 1.*** Proposed scheme for the acylation mechanism of *Burkholderia cepacia* lipase by acetic esters of primary alcohols. Only the main elements of the catalytic triad (Ser87, His286, Asp264), the oxyanion hole (NH of main chain Leu17 and Gln88), and the (R/S)-2-methyl-3-phenylpropyl (MPP) acetate ester are shown.

nism[20] and outlined by two consecutive stages: an enzyme acylation and a subsequent deacylation. The steps of acylation are summarized in Figure 1 for the case of *Burkholderia cepacia* lipase (BCL). A nucleophilic attack by the catalytic serine (Ser87) on the carbonyl carbon of an ester molecule leads to the formation of an acyl-enzyme adduct. Concurrently, the alcohol moiety is released via a negatively charged tetrahedral intermediate (THI)[1,7] involving the substrate.

Two transition states mark the two steps of the acylation reaction (see Figure 1): TS1, regulated by the nucleophilic attack by Ser87 and by the ability of a nitrogen atom ($N_\epsilon$) of the catalytic histidine (His286) to accept the proton from the serine. TS2, in which the proton is transferred from His286 to the substrate, with the formation of the acyl-enzyme.

Although the catalytic roles of serine and histidine residues are unanimously accepted, a wide variety of mechanisms have been proposed to explain the role of the acidic catalytic residues (Asp264 in the case of BCL) such as the single proton-transfer mechanism,[21–23] also supported by theoretical studies,[24,25] the double proton-transfer, also known as the "charge-relay" mechanism,[26,27] and the "low-barrier hydrogen bond"[28] and "short ionic hydrogen bond" mechanisms.[29]

A fundamental role in the formation and stabilization of THI is played by the oxyanion hole (formed by Gln88 and Leu17 in the case of BCL), which stabilizes the negatively charged species through hydrogen bonds. It is interesting to note that, while the THI often cannot be detected with the usual substrates,[1,30] recent studies have unambiguously shown that it is a shallow local minimum, not exceeding 3 kcal/mol in most cases.[31,32]

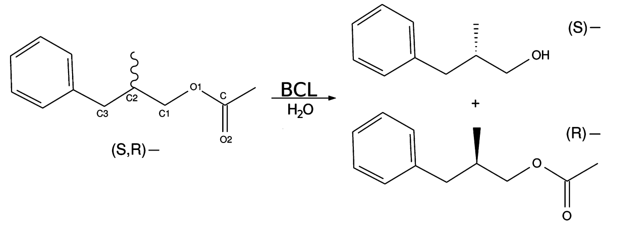


***Figure 2.*** Hydrolysis reaction of (S)-MPP acetate catalyzed by *Burkholderia cepacia* lipase.

***Table 1.*** Enantioselectivity (*E*) and Available Kinetic Constants for BCL-Catalyzed Hydrolysis of (R)- and (S)-MPP Esters

| substrate | $E^a$ | $k_{cat}$ (min $^{-1}$) | $K_M$ (mM) |
|---|---|---|---|
| (S)-MPP acetate[37] | 16 ± n.d | 4.7 ± 2.1 | 28 ± 1 |
| (R)-MPP acetate | | 3.2 ± 0.4 | 300 ± 140 |
| (S)-MPP butanoate[35] | 130 ± 30 | | |
| (R)-MPP butanoate | | | |
| (S)-MPP heptanoate[38] | ≥190 ± 30 | 0.4 ± 0.1 | 4 ± 1 |
| (R)-MPP heptanoate | | 0.004 ± 0.001 | 3 ± 1 |

[a] Enantioselectivity $E = (k_{cat}/K_M)_S/(k_{cat}/K_M)_R$.

The deacylation stage of the catalytic reaction, not considered in this study, proceeds via the nucleophilic addition of a lytic species (water or alcohol) to the carbonyl carbon of the acyl-enzyme. A subsequent proton transfer from the lytic species to the histidine residue[33] leads to the formation of the final product and restores the catalytic activity of the lipase.

Although all lipases exhibit enantioselectivity toward esters of secondary alcohols, *Burkholderia cepacia* lipase (BCL), formerly known as *Pseudomonas cepacia* lipase (PCL), is also able to react enantioselectively with several esters of primary alcohols.[14,34,35] In particular, BCL catalyzes the enantioselective hydrolysis of some esters of 2-methyl-3-phenyl-propanol (MPP), which is an important precursor in pharmaceutical synthesis (see Figure 2 and Table 1).[17,18] The experimental data shown in Table 1 underline that in these reactions BCL favors the (S)-enantiomers with varying enantioselectivity (*E*) values.[35–38]

The experimental data reported in Table 1 reveal that BCL enantioselectivity toward MPP esters is mainly determined by the chirality of the substrates' alcohol moiety: (S)-enantiomers are hydrolyzed faster than (R)-enantiomers. The acyl chain length comes into play as a modulating parameter, which strongly influences the measured kinetic values.

In the case of acetate, for instance, the enantioselectivity value is unambiguously due to the difference between the $K_M$ constants for the (S)- and (R)-enantiomers. This suggests small differences in binding strength are at the origin of the pristine enzyme enantioselectivity for short acyl chain lenghts.

For the heptanoate case, the similar $K_M$ values entail that there are no major differences in the binding strength, therefore pinpointing the enantioselectivity of BCL for this substrate to the large difference in the $k_{cat}$ values.

In the case of butanoate, finally, there is a substantial lack of kinetic data. However, the *E* value, similar in magnitude to the one of the heptanoate ester, endorses the conjecture that the butanoate kinetic data might show a similar trend to that of the heptanoate data.

Because of the industrial relevance of MPP, numerous experimental and theoretical studies have been performed to explain and improve the enantioselectivity of BCL toward racemic esters of this substrate.[34–44]

In particular, Mezzetti et al.[38] have recently resolved the X-ray structures of two phosphonate transition-state analogs of MPP heptanoate (hexylphosphonic acid (R/S)-2-methyl-3-phenylpropyl ester) bound to BCL, which contain the (R)- and (S)-enantiomers of the alcohol. In the two structures, the analogs bound to BCL in a similar manner, with the phenyl group of the alcohol pointing toward the solvent. The enantiomers adopt a so-called "mirror-image orientation" in which the methyl substituent (−CH3), the large substituent (Phe−CH2−) at the alcohol stereocenter, and the oxygen atom of the alcohol moiety (−O1−) are accommodated in "a similar position" in their respective complexes. As a consequence of this accommodation, due to stereochemistry requirements, the only hydrogen at the stereocenter has to "point in opposite directions".[38] In the discussion of their data,[38] Mezzetti et al. first remark that the relative orientations of the enantiomers in the X-ray structures differ significantly from all the predictions by previous modeling studies, focused on the simulation of enantiomer recognition.[37,39,41,43] Afterward, they try to rationalize what they call "an inconsistence": the existence of almost superposable bound phosphonate structures, mimicking the (R)- and (S)-tetrahedral intermediates having "a similar $K_M$ value [...] but 100-fold different $k_{cat}$ values" in favor of the (S)-enantiomer. Mezzetti et al. hypothesize, that BCL enantioselectivity, instead of stemming from a discrimination in the transition state, might rely on the possibility of the slow (R)-enantiomer binding to the enzyme in both a *productive* and a *nonproductive* way. Only productive binding would lead to catalysis. The fast (S)-enantiomer, on the other hand, is supposed to bind only in a productive way. However, while the authors propose this explanation for the enantioselectivity of BCL, they also notice that other possibilities, such as the lack of key interactions of the (R)-enantiomer within the active site of the enzyme, cannot be excluded a priori as possible explanations for the different observed enantioselectivity. Mezzetti et al. conclude their discussion stating that "given the subtlety of the interactions, it may be difficult to rationally predict substrate modifications or lipase mutations that would increase the enantioselectivity".

According to the insights from all the above-mentioned studies, the detailed mechanism and dynamics of the enantioselective biocatalysis by BCL is still an open question. Assessing the relative importance of chirality and acyl chain length in determining the experimental differences shown in Table 1 would certainly be an outstanding goal. Nonetheless, the complexity of the catalytic data is such that only a systematic study may result in a clear rationalization of $k_{\mathrm{cat}}$ and $K_M$ values. Recently, it has been underlined that attempts to estimate small $K_M$ differences with state of the art molecular dynamics based approaches might be unsuccessful,[45] consistent with the previous molecular modeling studies that failed in determining the relative orientations of enantiomers into the BCL active site.[37,39,41,43] Instead, molecular dynamics simulations have proven to be successful in many cases of $k_{\mathrm{cat}}$ prediction/rationalization.[45]

Driven by these motivations, it seemed critical to us facing the investigation of the catalytic mechanism of BCL toward primary alcohols, keeping the effect of the acyl chain length on enantioselectivity out. Accordingly, we simulated the acylation reaction of BCL on both enantiomers of MPP acetate by performing molecular dynamics (MD) simulations using a mixed quantum mechanical/molecular mechanical (QM/MM) approach. In particular, the free energy surfaces of the acylation reaction were reconstructed for both enantiomers using the metadynamics method,[46,47] allowing simultaneous analysis of the dynamical and structural aspects of the catalytic process during the 90 ps of MD performed for each enantiomer.

In this study, we observed that the reaction for (S)-MPP acetate proceeds in the expected way (see Figure 1), whereas the (R)-MPP substrate undergoes the reaction through a peculiar rearrangement of the active site: due to steric hindrance, the (R)-enantiomer induces a conformational change in the catalytic site, reorganizing a new triad (Ser87, His286, Glu289) instead of the native one (Ser87, His286, Asp264). As a consequence of the different reaction mechanisms, the two enantiomers show similar activation free energies ($\Delta G^{\ddagger}_{\mathrm{S}} = 20.5 \pm 2$ kcal/mol for (S)-MPP and $\Delta G^{\ddagger}_{\mathrm{R}} = 17.3 \pm 2$ kcal/mol for (R)-MPP) and consequently similar $k_{\mathrm{cat}}$ values. The MPP acetate enantiomers are therefore discriminated only at the level of the binding strength, since both enantiomers can then find similar energetically favorable reaction pathways.

On the contrary, a comparative analysis of our results with the experimental X-ray investigations on transition state analogs of heptanoate suggests that both enantiomers of that substrate have to follow an identical reaction pathway. The bulky acyl chain of heptanoate is expected to prevent the Val266 movement necessary for the His286 flip rotation. Without histidine reorganization, the enzyme is forced to work in the native form, destabilizing the acylation reaction for the (R)-enantiomer. This destabilization, due to steric hindrance, might be the source of the large difference between the kinetic constants $k_{\mathrm{cat}}$ for (R)- and (S)-MPP heptanoate.

## 2. Methods and Computational Details

**2.1. System Setup.** System setup and classical MD were performed with the NAMD (v.2.6)[48] and VMD (v.1.8.6)[49] software packages. The standard AMBER[50] forcefield was used for the protein, ligands, and counterions. The water was modeled using the TIP3P[51] forcefield. Atom types as well as bonded and nonbonded parameters were assigned to atoms by analogy or through interpolation from those already present in the forcefield. To calculate partial atomic charges, an "elongate" conformation of (S)-MPP acetate was optimized using the *ab initio* quantum chemistry program GAMESS[52] at the HF/6-31G* level of theory. Consequently, a set of atom-centered charges were obtained by applying the ESP methodology as implemented in the ELPOT and PDC modules of GAMESS. Charges on equivalent atoms were equalized by averaging. The same charge values were used for the (R)-MPP acetate.

The initial substrate conformations were derived from the phosphonate inhibitor (hexylphosphonic acid (R/S)-2-methyl-3-phenylpropyl) of the BCL X-ray structure deposited in the Protein Data Bank (PDB) with accession code 1YS1 and 1YS2 for the (R)-MPP and the (S)-MPP enantiomers, respectively. The substrates were then superimposed on the isomorphous crystal structure of the activated form of the BCL[53] (PDB code 3LIP). The hexyl chain of the phosphonate esters was replaced by a methyl group, and the phosphorus atom was replaced by a carbon atom. Hydrogen atoms were added to the system using the VMD tools. Histidines were uncharged and protonated according the most plausible hydrogen-bonding pattern in the structure. Aspartic and glutamic acids were negatively charged, and arginines and lysines were positively charged. A geometry relaxation of the substrate molecules and catalytic residues with the rest of protein heavy atoms fixed to the crystallographic positions and without water was performed with 5000 steps of conjugate gradient geometry optimization through a NAMD minimization tool.

Each complex was then surrounded by a periodic box of TIP3P[51] water molecules, and 4 $Na^+$ ions were added using the VMD tools,[49] to guarantee neutrality. The same spatial arrangement of ions was adopted for both systems, placing the ions at a distance greater than 25 Å from the catalytic serine to minimize the charge effect. The total number of water molecules was 7094 in an initial rectangular box with dimensions of 63 × 70 × 62 Å.

The simulations were conducted using periodic boundary conditions and the long-range part of the electrostatics was treated with the Particle-Mesh-Ewald (PME) method,[54] with a grid size of 64 × 70 × 62. The cutoff distance for nonbonded interactions was set to 11 Å, and a switch function was applied to smooth interactions between 10 and 11 Å. The scaling factor used in NAMD for 1−4 intramolecular Coulomb interactions was set to 0.8333, which is the inverse of the standard scaling factor value used in the input AMBER file (SCEE = 1.2). The r-RESPA multiple time step method[55] was employed with 1 fs for bonded, 2 fs for the short-range part of the nonbonded, and 4 fs for the long-range part of the electrostatic forces.[48]

All simulations were conducted in the NPT ensemble. The temperature was set to 300 K and controlled via Langevin thermostat.[56] The pressure was set to 1 atm and controlled via isotropic Langevin piston manostat.[57]

The systems were submitted to 600 ps of MD simulation. During the first 200 ps, the proteins' heavy atoms and $Na^+$ ions were harmonically restrained with a force constant of 10 kcal/(mol Å$^2$). Subsequently, 200 ps of dynamics were performed with the force constant set to 5 kcal/(mol Å$^2$), and finally 200 ps with the force constant set to 1 kcal/(mol Å$^2$). This allowed the equilibration of the solvent and the proper readjustment of the cell volume without disruption of the ligand conformation, or of the lipase structure that preserved the active open form.[7] The final structures were then used as starting point for QM/MM simulations. To ensure the validity of the structures obtained, a subsequent run was continued for a further 100 ps without restraints and with velocities redistributed according to a Boltzmann distribution. The mean values of the protein heavy atoms RMSD calculated with respect to the initial conformation were 0.87 and 0.92 Å for the (R)-MPP and (S)-MPP systems, respectively (RMSD time evolutions are available in the Supporting Information).

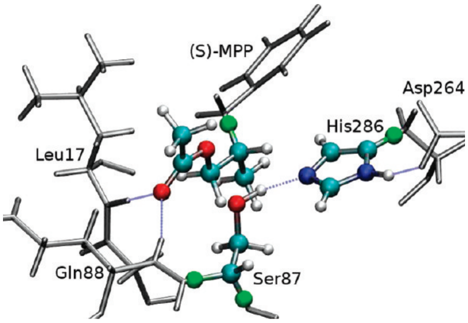


*Figure 3.* Main residues of the catalytic site and (S)-MPP acetate substrate. The atoms comprising the QM region that were used in this study are shown in ball-and-stick representation. The link hydrogen atoms are highlighted in green; the MM region, in gray, is shown in tube representation.

**2.2. QM/MM System.** In order to obtain an accurate description of the chemical processes involved in the reaction mechanism while reducing the computational cost inherent in *ab initio* calculations, a QM/MM scheme was chosen to model the reactive site.

The QM/MM driver[58,59] is based on the QM program QUICKSTEP[60,61] and the MM driver FIST, which are both part of the freely available CP2K package.[62] The quantum part is treated at the density functional theory (DFT) level. This region consists of the substrate, with the exception of the large substituent (Phe−CH2−), the Ser87 side chain with inclusion of the $C_\alpha$ backbone atom (at a two atoms distance from the attacking oxygen), and the imidazole ring of His286, adding up to a total number of 35 atoms. This set contains all the atoms directly involved in the reaction or whose stabilization of the reaction intermediates cannot be described uniquely with electrostatic effects. Therefore, the only atoms which have been excluded from the quantum region are the hydrogens of the oxyanion hole and side chain of Asp264 and Glu289 near the His286. The exclusion of these atoms from the quantum region is not crucial due to the evidence[24,25,33,63] that they have mainly an electrostatic stabilization role.

The boundary between the QM and the MM regions ($C_\gamma$−$C_\beta$ of His286, $C_\alpha$−CO and $C_\alpha$−NH of Ser87, and C2−C3 of MPP) were saturated by link hydrogen atoms (see Figure 3). In agreement with the IMOMM link scheme,[64] the scaling factor projecting the forces on the capping hydrogen was refined to maintain the QM/MM bond distances at the same values of the forcefield.

The remaining part of the system, including the water molecules and counterions, has been modeled at the classical level with the AMBER forcefield, explicitly taking into account the steric and electrostatic effects of the substrate, the enzyme, and the solvent.

A triple-ζ valence basis set with two sets of polarization functions, TZV2P,[65] and an auxiliary plane-wave basis set with a density cutoff of 280 Ry were used to describe the wave function and the electronic density. Dual space pseudopotentials[66,67] were used for describing core electrons and nuclei. We used the gradient-corrected Becke exchange[68] and the Lee, Parr, and Yang correlation functional (BLYP).[69] Energies were tested for convergence with respect to the wave function gradient ($5 \times 10^{-7}$ Hartree) and cell size, which was required to be no smaller than 16.0 Å (cubic box) to achieve a correct decoupling between the periodic images.[70]

The QM regions were first minimized by keeping the entire MM subsystem frozen. Subsequently, a complete minimization was performed over the entire system, employing a conjugate gradient method as implemented in CP2K. Root-mean-square (RMS) values of 0.005 hartree·Bohr$^{-1}$ for force and 0.005 Bohr for positions were selected as convergence criteria. The systems were then subject to a brief equilibration by means of a QM/MM MD simulation in the NVT ensemble for 2.5 ps. The temperature was set to 300 K, and each degree of freedom was controlled via a Nosé–Hoover thermostat[71,72] with a time constant of 50 fs. A single integration time step of 0.50 fs was used.

**2.3. Metadynamics.** The choice of collective variables (CV) is crucial in metadynamics[46,47] for its successful application. Looking at Figure 1, the fundamental geometrical variables describing the acylation reaction can be identified as the distances between the atoms involved in the nucleophilic attack, the subsequent release of the alcohol's moiety, and the hydrogen transfer.

During the nucleophilic attack, the $r_{C-O}$ variable, defined as the distance between the carbonyl carbon atom of the ester group and the serine oxygen atom, gradually decreases until the formation of the acyl-enzyme complex. In constrast, the $r_{C-O1}$ variable, defined as the distance between the carbonyl carbon atom of the ester and the alcohol oxygen atom, gradually increases during the nucleophilic attack and the subsequent release of the alcohol moiety. The variable $r_{H-N_\epsilon}$, defined as the distance between the serine hydrogen and the histidine nitrogen, traces the proton transfer that occurs during the reaction.

To reduce the complexity while maintaining an accurate description of the system, 2 CVs were chosen from the three variables mentioned and defined as $CV_a(r) = r_{H-N_\epsilon}$ and $CV_b(r) = (r_{C-O1} - r_{C-O})$, the difference between the bond distances $r_{C-O1}$ and $r_{C-O}$, spanning a two-dimensional subspace of the free energy surface reaction. The metadynamics runs were performed using Gaussian-shaped potential hills with a height of $3.0 \times 10^{-3}$ Hartree and a width of 0.1 Bohr.

The hills were spawned at intervals of 20 fs of QM/MM MD. To restrict the surface of exploration, an upper limit in the $CV_a(r)$ was imposed with activation of a quadratic wall positioned at 2.20 Å, with a quadratic potential constant of 30.0 kcal/(mol Å$^2$), whereas $CV_b(r)$ was delimited by two quadratic walls positioned at −1.5 and 1.5 Å, with a quadratic potential constant of 20.0 kcal/(mol Å$^2$). QM/MM metadynamics were conducted in the NVT ensemble. The temperature was set to 300 K, and each degree of freedom was controlled via a Nosé-Hoover thermostat[71,72] with a time constant of 50 fs. Temperature stability was monitored along the metadynamics runs (see the Supporting Information). A single integration time step of 0.5 fs was used. The runs were protracted for about 90 ps for both systems until they showed a free diffusivity along the CVs. These convergence criteria were chosen in agreement with the guidelines published in a recent paper directed to assess the accuracy of metadynamics.[73] Trajectories were saved every 20 steps (10 fs time interval) of metadynamics for subsequent analysis. The long QM/MM MD simulation times guaranteed an extensive sampling of the configurational space, important for providing meaningful determination of the energetics for enzymatic reactions.[74]

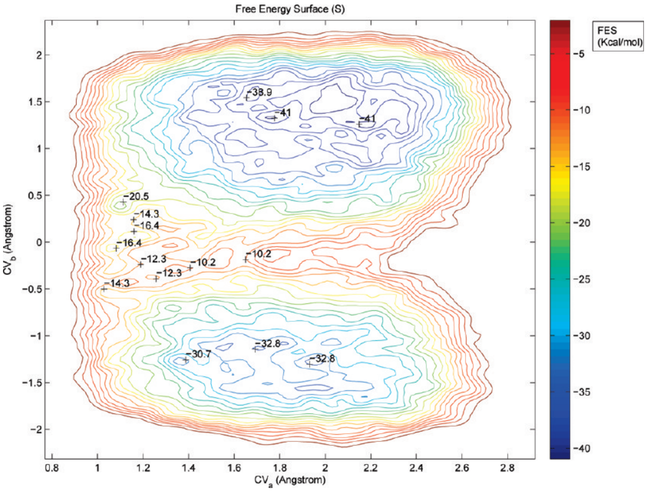


***Figure 4.*** Acylation reaction free energy surface of the (S)-MPP system reconstructed using metadynamics as a function of two CVs, specifically, $CV_a(r) = r_{H-N_\epsilon}$ and $CV_b(r) = (r_{C-O1} - r_{C-O})$. Energy is in kcal/mol; the CV values are expressed in Å.

## 3. Results

To elucidate catalytic mechanism of BCL at the molecular level, we reconstructed the free energy surfaces (FESs) for the enzyme acylation reaction by MPP acetate enantiomers according to two defined CVs, $CV_a(r) = r_{H-N_\epsilon}$ and $CV_b(r) = (r_{C-O1} - r_{C-O})$, using metadynamics and a QM/MM computational framework.

The results of our calculations are given in Figures 4 and 5 for (S)-MPP acetate and (R)-MPP acetate, respectively. The two FESs show similar contours, and in both cases, two broad minima can be recognized corresponding to the enzyme–substrate complex (ES) and to the enzyme–product complex (EP). The ES region spans from 0.75 to 2.5 Å for $CV_a(r)$ and negative values for $CV_b(r)$. The EP region corresponds to values from 0.75 to 2.5 Å for $CV_a(r)$ and positive values for $CV_b(r)$.

In Figure 6, the time evolution of $CV_b(r)$ is displayed together with the times at which the acylation reaction and the reverse reaction have occurred. The observation of the "retro" reaction restoring the reagent species highlights the early achievement of free diffusivity conditions along the CVs. According to the description of the reaction profile,

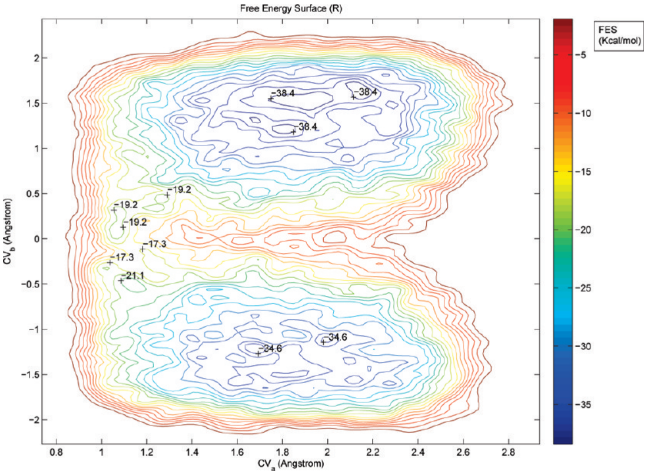


***Figure 5.*** Acylation reaction free energy surface of the (R)-MPP system reconstructed using metadynamics as a function of two CVs, specifically, $\mathrm{CV_a}(r) = r_{\mathrm{H-N}_\varepsilon}$ and $\mathrm{CV_b}(r) = (r_{\mathrm{C-O1}} - r_{\mathrm{C-O}})$. Energy is in kcal/mol; the CV values are expressed in Å.

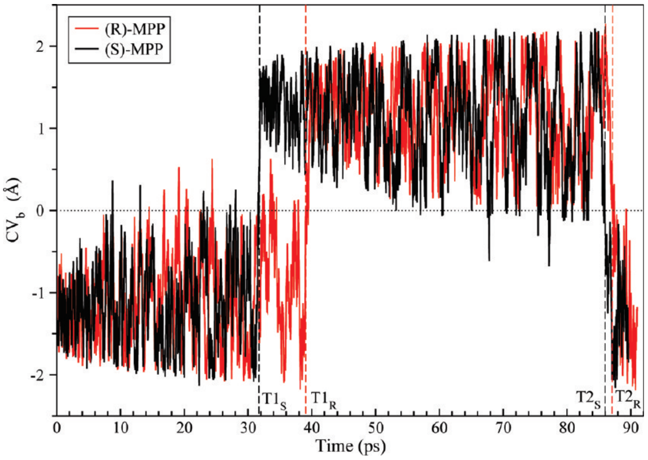


***Figure 6.*** Time evolution of $\mathrm{CV_b}(r)$ during the metadynamics run. The trend of the $\mathrm{CV_b}(r)$ for (R)-MPP is shown in red, whereas the trend of the $\mathrm{CV_b}(r)$ for (S)-MPP is given in black. Vertical lines give the time at which the reaction takes place. The labels $\mathrm{T1_S}$ and $\mathrm{T1_R}$ refer to the acylation reaction for the (S)- and (R)-enantiomers, respectively. $\mathrm{T2_S}$ and $\mathrm{T2_R}$ refer to the reverse reaction.

it can be observed that the reaction occurs after 32 ps in the case of the (S)-enantiomer and after 39 ps in the case of the (R)-enantiomer, that is, when $\mathrm{CV_b}(r)$ changes from negative to positive values.

Both acylation reactions are exothermic, and the total free energy changes, extrapolated from the values reported in Figures 4 and 5, are about $\Delta G = -8 \pm 2$ kcal/mol for (S)-MPP and $\Delta G = -3 \pm 2$ kcal/mol for (R)-MPP. As proposed in Figure 1, both reactions proceed from the ES complex to the TS1 species through nucleophilic attack and proton transfer. The TS1 structures for both enantiomers can be localized on the FES near a value of 1.3 Å for $\mathrm{CV_a}(r)$ and −0.5 to 0.0 Å for $\mathrm{CV_b}(r)$.

In the case of (S)-MPP acetate, the acylation reaction proceeds from TS1 to the proposed THI: in Figure 4, near a value of about 0 Å for $\mathrm{CV_b}(r)$ and about 1.1 Å for $\mathrm{CV_a}(r)$, it is possible to identify a shallow basin which reflects the presence of a transient species, possibly to be identified as the THI. The resolution of the metadynamics run, however, proportional to the height of the Gaussian functions ($3.0 \times 10^{-3}$ Hartree), is of about the same magnitude as the observed stabilization energy for the supposed THI local minimum. Due to this remark and since detailed inspection of trajectories did not succeed in identifying a proper geometrical characterization of the supposed THI, one cannot discard the possibility that such a minimum is only an aberration due to metadynamics. Finally, after the shallow basin, the reaction goes toward the presumed TS2, with the consecutive release of the (S)-alcohol.

In the case of (R)-MPP, the FES region corresponding to TS1, THI, and TS2 is extremely flat, and the tetrahedral intermediate is not as readily identifiable (see Figure 5) as in the (S)-MPP case. In the EP basin, the reaction proceeds toward the release of the alcohol moiety in a similar way as it does with (S)-MPP.

Although Figures 4 and 5 display some differences in the contours of the FESs, notably in the EP basin, both surfaces exhibit extremely similar topological features. A quantitative analysis reveals that the free energy barrier in going from the ES basin to the first transition state (TS1) is $\Delta G^{\ddagger}_{\mathrm{S}} = 20.5 \pm 2$ kcal/mol for (S)-MPP and $\Delta G^{\ddagger}_{\mathrm{R}} = 17.3 \pm 2$ kcal/mol for (R)-MPP. The reverse reaction (from the EP basin to the transition state TS2) occurs with a free energy barrier of about 29 ± 2 and 21 ± 2 kcal/mol for (S)-MPP and (R)-MPP, respectively. As a whole, the difference in the total free energy change for the MPP acetate ester is $\Delta\Delta G_{\mathrm{S-R}} = -5 \pm 2$ kcal/mol in favor of the release of the (S)-MPP alcohol moiety.

The large magnitude of the barriers involved in the forward and backward reaction paths addresses both enantioselective hydrolyses of the acetate esters as thermodynamically controlled reactions. In fact, at least for the (R)-enantiomer, the barriers of both forward and backward reactions are comparable, within computational errors. This casts serious doubts on the possibility to rationalize quantitatively the data reported in Table 1 simply applying the conventional Michaelis−Menten mechanism, which correlates the kinetic constant $k_{\mathrm{cat}}$ to the forward reaction only and assumes the acylation step to be irreversible.[75] Therefore, the atomistic aspects involved in the acylation reaction were analyzed to properly understand the significance of kinetic and calculated thermodynamic data.

To investigate the reaction mechanism, some geometrical variables describing the motion of the catalytic triad along the metadynamics trajectory were monitored. In particular, the dihedral angles $\chi_1$ (along the $\mathrm{C}_\alpha-\mathrm{C}_\beta$ bond) and $\chi_2$ (along the $\mathrm{C}_\beta-\mathrm{C}_\gamma$ bond) describe the dynamic behavior of the imidazole ring of His286 during the reaction. Their time evolution is reported in Figures 7 and 8, respectively.

In the case of the (S)-MPP acetate, the time evolution of $\chi_1$ values shows a step at about 30 ps corresponding to a main movement (see Figure 7). This can be described as a sort of "pivoting" motion of the His286 imidazole ring, which drives the hydrogen proton transfer from the Ser87 residue

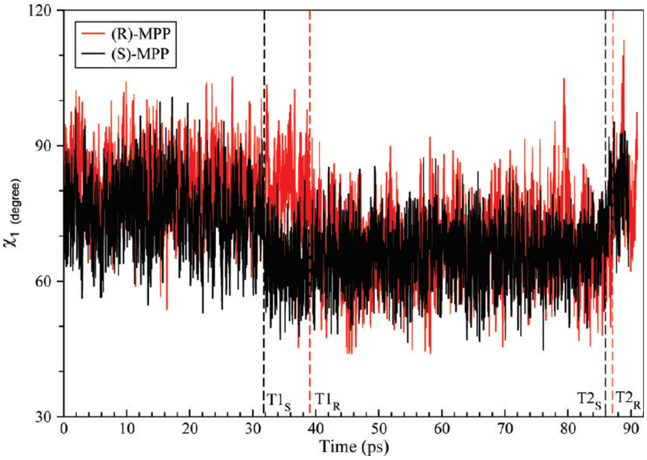


***Figure 7.*** Time evolution of the $\chi_1$ dihedral angle along the $C_\alpha-C_\beta$ bond of His286 during the metadynamics run. Red shows the course of $\chi_1$ for (R)-MPP, black that of $\chi_1$ for (S)-MPP. Vertical lines indicate the time at which the reaction takes place, as described in the caption of Figure 6.

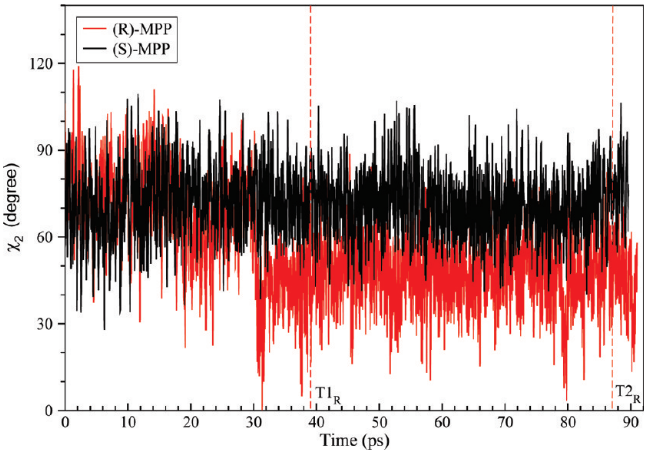


***Figure 8.*** Time evolution of the $\chi_2$ dihedral angle along the $C_\beta-C_\gamma$ bond of His286 during the metadynamics run. Red depicts the course of $\chi_2$ for (R)-MPP, black that of $\chi_2$ for (S)-MPP. Vertical lines indicate the time at which the reaction takes place, as described in the caption of Figure 6.

to the oxygen atom of the alcohol moiety. As in contrast to the dihedral angle $\chi_1$, the time evolution of the dihedral angle $\chi_2$, displayed in Figure 8, reveals only fluctuations of the imidazole ring plane.

Figure 9 displays the superposition of representative snapshots of the metadynamics trajectory of (S)-MPP acetate. The structures identify a progression from an initial to a final reaction state passing through an "intermediate" species. As shown in Figure 9, the $-CH_2-$ group of the large substituent of the (S)-MPP enantiomer points toward Leu17. This avoids any steric hindrance between His286 and the alcohol moiety, allowing the pivoting of His286 and the reaction to occur smoothly.

In the case of the (R)-MPP acetate, the time evolution of the $\chi_1$ values shows a similar trend with respect to (S)-MPP, while a peculiar rotation of about 30−40° of the $\chi_2$ dihedral angle is observed (red in Figure 8), which occurs at about 30 ps, that is, 10 ps before the reaction takes place. The motion corresponds to a rotation or "flip" of the imidazole ring. In Figure 10, the superposition of two representative structures extracted from metadynamics runs, sampling the catalytic environment before and after the histidine flip, is shown. From the two snapshots of Figure 10, we can reconstruct the His286 rearrangement path, which is induced by the mechanical effect of a clash between the (R)-alcohol moiety and the imidazole ring. Interesting enough, it is possible to observe that, after the flip has occurred, the alcohol moiety can get closer to the catalytic triad, with a

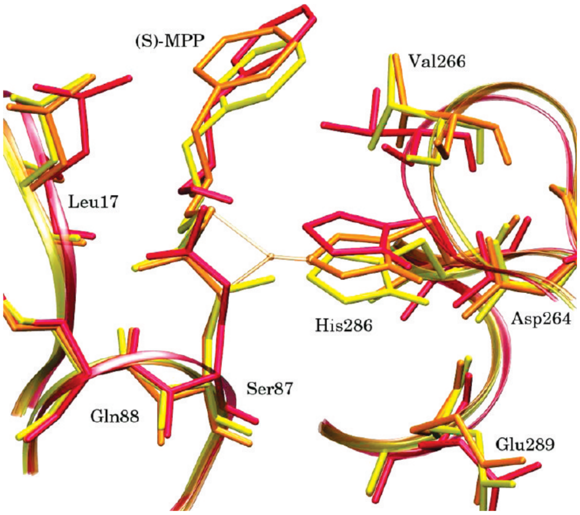


***Figure 9.*** Superposition of representative snapshots of the metadynamics trajectory showing the evolution of the positions of the catalytic triad during the acylation reaction for (S)-MPP ester. The color code (yellow → orange → red) corresponds to structures at 31.57, 31.72, and 54.43 ps of simulation, respectively.

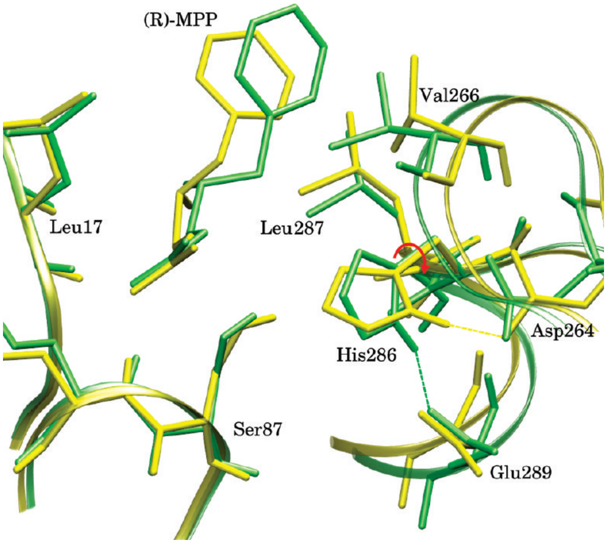


***Figure 10.*** Catalytic triad residues. (R)-MPP ester enantiomer and main residues surrounding the catalytic triad are displayed in stick representation. Only polar hydrogens are shown. A representative snapshot of the initial conformation is depicted in yellow. A representative snapshot of the metadynamics run (40 ps) after rotation of the catalytic histidine is depicted in green. The hydrogen bond between His286 and Glu289 is depicted as a dashed green line, while the hydrogen bond between His286 and Asp264 is depicted as a dashed yellow line. The $\chi_2$ dihedral angle is depicted as a red arrow.

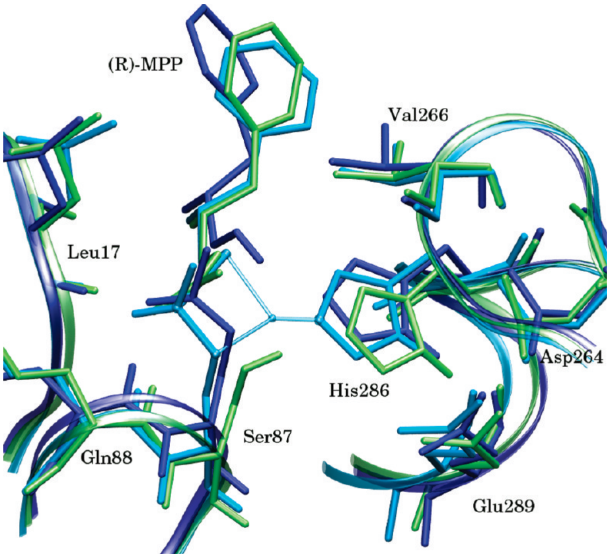


***Figure 11.*** Superposition of representative snapshots of the metadynamics trajectory showing the evolution of the positions of the catalytic triad during the acylation reaction for (R)-MPP ester. The color code (green → cyan → blue) corresponds to structures at 35.34, 39.65, and 67.69 ps of simulation, respectively. The progression is shown from an initial to a final reaction state passing through an "intermediate" species represented by a selected snapshot of the metadynamics trajectory during the hydrogen transfer process. Even if the $-CH_2-$ group of the large substituent of the (R)-MPP enantiomer points toward His286, the reaction proceeds (after flip motion) as in the case of the (S)-enantiomer.

portion of the large substituent (Phe−CH2−) inserted more deeply into a cleft formed by the side chain of Leu287 and His286, also known as the "His gap".[76,77] Therefore, the His286 ring rotation increases the size of the narrow "His gap" cleft, facilitating the approach of the (R)-MPP alcohol toward the catalytic triad. It is important to observe that the flip movement is permitted by a preparatory rearrangement of the side chain of Val266, which rotates by about 120° with respect to the initial position. As shown in Figure 10, after this rotation, one of the Val266 methyl groups no longer points toward the histidine imidazole, allowing for a greater mobility of this residue.

During the histidine flip, the hydrogen bond between the catalytic triad residues His286 and Asp264 breaks, and another hydrogen bond forms between the rotated His286 and Glu289 to restore an alternative catalytic triad (Ser87, His286, and Glu289). Because of the current choice of CVs, the transition state associated with the catalytic triad reorganization has not been detected, as it lays in a space orthogonal to the one we used for exploring the free energy surface. However, after such a reorganization of the catalytic triad, the acylation reaction of the (R)-enantiomer proceeds as in the case of the (S)-enantiomer and can be described with the selected CVs without a loss of accuracy. Figure 11 shows the superposition of representative snapshots of the metadynamics trajectory after the flip motion, showing the time evolution of the (R)-enantiomer and of the residues surrounding the substrate, including the alternative catalytic triad, during the reaction.

Figures 9 and 11 allow the reconstruction of a path for the acylation reaction, which is consistent with the serine protease mechanism suggested by Radisky et al.[78] and by Fuhrmann et al.,[29] on the basis of experimental evidence. Moreover, Figures 9 and 11 highlight the different "adaptation" of the enzymatic catalytic triad to the two enantiomers. In the (R)-MPP system, the preliminary histidine flip motion was necessary for the occurrence of the reaction; on the contrary, the acylation reaction of the (S)-MPP ester occurred without any disruption or rearrangement of the catalytic triad.

To the best of our knowledge, this is the first time that clear evidence is provided, at the atomistic level, of the possibility for BCL to shift toward a secondary catalytic triad by preliminary histidine flip motion.

It is worth noting that, as demonstrated by the computed FESs, (i) BCL retains similar catalytic activity using the secondary triad Ser87, His286, and Glu289 and (ii) the reorganization of the enzyme environment is associated with a destabilization of the acyl-enzyme adduct, which is reflected in the smaller free energy difference calculated for the acylation reaction of (R)-MPP acetate.

The flip motion has been observed because BCL contains two alternative acidic residues able to participate in the formation of the catalytic triad: Asp264 and Glu289. A closely related lipase, *Pseudomonas glumae*, exhibits the same characteristic arrangement of acidic residues, Asp263 and Glu288, equivalent to Asp264 and Glu289 of BCL.[79,80] In that case, mutation of Asp263 into alanine yielded a lipase with 25% of the original activity.[81] This experimental observation strengthens the point that, similarly to *Pseudomonas glumae*, the Glu289 residue of BCL can also serve as an alternative proton acceptor.

The different catalytic mechanisms for the (S)- and (R)-enantiomers suggest a way to increase further BCL enantioselectivity by mutation of Glu289 into an aprotic residue such as alanine. In fact, the lack of an additional donor for the formation of the secondary catalytic triad would render the standard catalytic triad the only one accessible for the conversion of the (R)-enantiomer, with a large steric hindrance. This would destabilize the reaction pathway, providing a clean way to deplete the catalytic activity for one of the two enantiomers.

Other structural adjustments of the enzyme environment can be readily observed by root mean-square deviation (RMSD) and root mean-square fluctuation (RMSF, i.e. standard deviation) values of individual side chain residues within 6 Å of the catalytic triad, calculated with respect to the starting position of the metadynamics run. RMSDs and RMSFs for selected residues, time averaged over the simulation, are reported in Figures 12 and 13 for both enantiomers.

The residues with the highest RMSD are Glu289, Phe119, and Val266. Glu289 shows a high RMSD and RMSF only in the case of the (S)-MPP because it is not involved in any hydrogen bond within the catalytic triad. The mobility of this residue, on the contrary, is reduced in the (R)-MPP system due to the hydrogen bond formation with His286 in the secondary catalytic triad. Phe119 and Val266 are adjacent residues and form part of a hydrophobic pocket, which

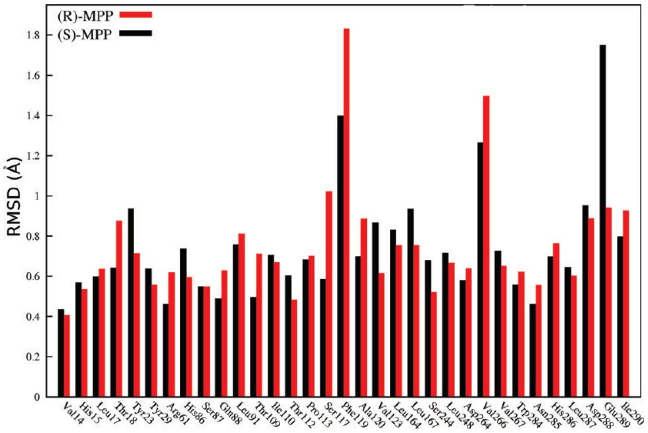


***Figure 12.*** RMSD of residue side chain atoms. Only residues within 6 Å of the catalytic triad are reported. Red bars correspond to the (R)-MPP enantiomer; black bars correspond to the (S)-MPP enantiomer.

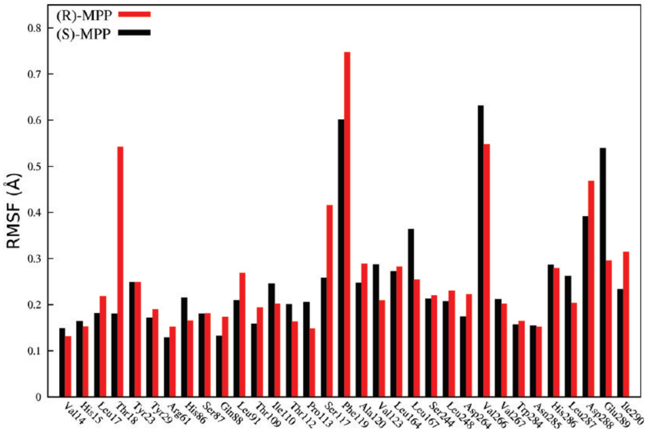


***Figure 13.*** RMSF of residue side chain atoms. Only residues within 6 Å of the catalytic triad are reported. Red bars correspond to the (R)-MPP enantiomer; black bars correspond to the (S)-MPP enantiomer.

comprises also Pro113, Leu164, Leu167, and Val267, also known as the hydrophobic groove HA,[82] where the acyl chain of the esters is accommodated. Accordingly, the fluctuations of the Phe119 side chain have been related to the emptiness of the HA groove usually occupied by acyl chains. The movement of the Val266 residue deserves more attention: it was monitored by following the time evolution of the dihedral angle along the $C_\alpha-C_\beta$ bond of this residue and of the distance between His286 $C_{\delta 2}$ and Val266 $C_\beta$ atoms (see in Figure 14A and B). In the case of the (R)-MPP system, as discussed above, a rotational rearrangement of the Val266 side chain occurs at about 20 ps (see Figure 14A), before both the His286 flip and pivoting movements. In the case of the (S)-MPP system, a comparable side chain rotation takes place as well, however, as a consequence of the reaction progress. Figure 9 shows that this rotation for the (R)-enantiomer is promoted by the pivoting movement of His286 and allows the relaxation of this residue.

An insight on the role of the acyl chain in the acylation reaction has been obtained superimposing the previously described X-ray structures of the hexylphosphonic transition-state analogues[38] onto the “intermediate” snapshots of the

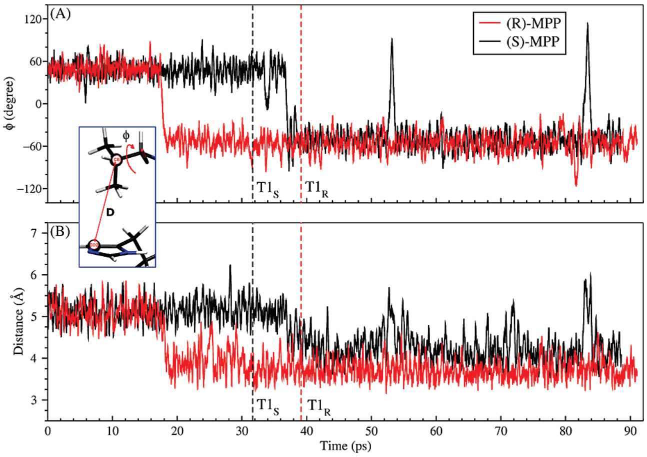


***Figure 14.*** Time evolution of the dihedral angle along the $C_\alpha-C_\beta$ bond of Val266 (A) and of the intermolecular distance between His286 $C_{\delta 2}$ and Val266 $C_\beta$ atoms (B) during the metadynamics runs. The trend of the values for (R)-MPP is shown in red, whereas the trend of the values for (S)-MPP is given in black. Vertical dashed lines highlight the time at which the acylation reaction takes place. Labels $T1_S$ and $T1_R$ refer to the acylation reaction for the (S)- and (R)-enantiomers, respectively. In the blue box, a representative snapshot of the initial Val266 and His286 relative orientation is depicted: a red line shows the monitored distance (*D*), while a red arrow shows the monitored dihedral angle (*ϕ*).

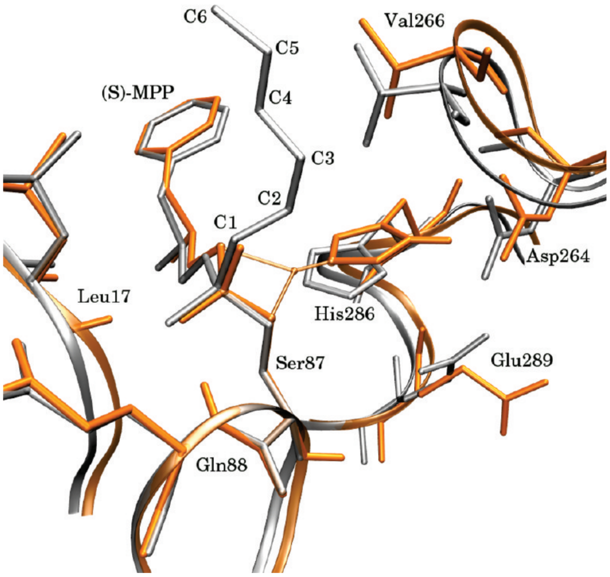


***Figure 15.*** Superposition of a molecular dynamics snapshot of the (S)-MPP “intermediate” species (orange) on the X-ray S-enantiomer phosphonate analogue structure (white).

(S)-MPP and (R)-MPP species (Figures 15 and 16). Although comparison with transition state analogues should be used with caution,[83] since here the catalytic histidine is not involved in the proton transfer and the system is in a relaxed conformation, the excellent overlap displayed in the figures indirectly confirms the accuracy of our computational setup.

As shown in Figures 15 and 16, the hexyl chains of both phosphonate esters extend into the hydrophobic groove HA. Both X-ray structures exhibit a favorable interaction between Val266 and the acyl chain, mainly because one of the Val266 methyl groups points toward the catalytic His286. This

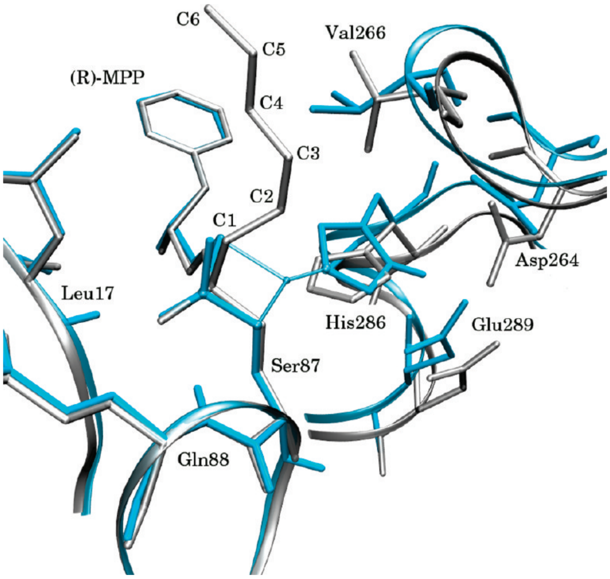


***Figure 16.*** Superposition of a molecular dynamics snapshot of the (R)-MPP "intermediate" species (cyan) on the X-ray structure of the (R)-enantiomer phosphonate analogue (white).

peculiar arrangement is necessary to avoid a remarkable steric hindrance between the rotated Val266 and the C3−C4 positions of the phosphonate acyl chain that would arise further to a hypothetical rearrangement of the catalytic triad (see Figure 16). Therefore, by choosing the acyl chain length appropriately, the rotation of the residue Val266 can be optimally controlled or even prevented, which supports the experimental evidence reported in Table 1 that acetate exhibits a moderate enantioselectivity and heptanoate has the largest enantioselectivity, whereas butanoate shows an intermediate value.

The results of this study offer an important interpretation of the kinetic data shown in Table 1. The similar values for $k_{cat}$ in the case of acetate compared to the different values of $k_{cat}$ for heptanoate can be justified with the finding that, in the case of acetate, the (R)- and (S)-enantiomers follow two different reaction pathways, using two different catalytic triads. Therefore, for the acetate, the enantioselectivity originates only at the level of the binding strength, since both enantiomers can find energetically favorable reaction pathways.

On the contrary, in the case of heptanoate, the reaction pathway is identical for (R)- and (S)-enantiomers. A possible rationale is that the accommodation of an heptanoil chain in the active site of BCL prevents the Val266 side chain movement so that the His286 flip rotation is hampered and the catalytic triad of the enzyme is "forced" to work in the native form. The blocked conformation thus creates a steric hindrance between the (R)-enantiomer and His286, which destabilizes the acylation reaction. This destabilization, which is present only for the (R)-enantiomer, is the source of the large difference between the two kinetic constants $k_{cat}$ for (S)- and (R)-MPP heptanoate.

## 4. Conclusion

Enantioselectivity is fundamental to the design of new and more efficient synthetic routes for modern drugs.

In this work, we report a study on the BCL-catalyzed hydrolysis of the acetic ester of (R/S)-2-methyl-3-phenylpropanol (MPP), which is an important precursor in pharmaceutical synthesis, using a QM/MM scheme based on DFT for treating the quantum region. Using metadynamics for both (S)- and (R)-enantiomers, we computed the free energy surfaces of the catalyzed reaction with respect to two collective variables that mapped the entire reaction path.

Our results show that the (R)-enantiomer cannot efficiently undergo the acylation reaction using the BCL native catalytic triad Ser87, His286, and Asp264, due to steric hindrance. This enantiomer, instead, can follow a different and fruitful reaction path exploiting an alternative triad based on Ser87, His286, and Glu289. Residue Glu289, therefore, due to its closeness to Asp264 and His286, plays a fundamental role as an alternative proton acceptor.

From our studies, we were also able to identify in residue Val266 one of the sources of the different stereospecifity shown by BCL toward substrates with different acyl chain lengths. A short acyl chain, in fact, allows mobility of Val266, which is essential to promote the flip motion of His286. On the contrary, longer acyl chains obstruct Val266 side chain rotation. A smaller mobility of this residue makes the rearrangement of His286 more difficult and, consequently, increases the stereospecifity of the catalyzed reaction.

The present work contributes, with state-of-the-art computer simulations, to the understanding of atomistic aspects of the catalytic triad in BCL when reacting with two different enantiomers of MPP and provides detailed information on how to regulate the enantioselectivity of this enzyme. In fact, we hypothesize that the mutation of Glu289 or Asp264 into an aprotic residue, or mutations designed to affect the mobility of either Val266 or the catalytic triad, would be possible strategies to regulate the stereoselectivity of the BCL lipase.

After the execution of this work, a paper concerning the control of BCL enantioselectivity by engineering the substrate accessibility channel appeared,[84] which gives experimental evidence of our theoretical insights on the essential contribution of Val266 to the enzyme enantiomeric preference. Lafaquière et al., working with (R/S)-2-chloro ethyl 2-bromophenylacetate, have in fact observed a reversal of the enantiopreference by mutation of Val266 into a "most compact glycine". According to our catalytic mechanism rationalization, annihilation of Val266 side chain hindrance is very much expected to affect the catalytic histidine mobility producing dramatic consequences on the enzyme enantioselectivity.

**Acknowledgment.** We thank the Swiss National Supercomputing Center (CSCS) for providing the computational resources used in this work on the IBM sp5 machine. L.B. gratefully acknowledges Prof. Jürg Hutter (University of Zurich Institute of Physical Chemistry) for his hospitality and the opportunity to collaborate with the CP2K development team. Project FIRB 2003 "Novel environment friendly biotransformations and optimization of industrial processes based on the use of enzymes and organometallic catalysts", protocol number RBNE 034XSW_005, of Ministero Italiano della Universitá e della Ricerca, is gratefully acknowleged.

**Supporting Information Available:** Further details about system setup, forcefield parameters, RMSD equilibration data and metadynamics configurations are provided in the Supporting Information. This information is available free of charge via the Internet at http://pubs.acs.org/.

## References

(1) Hedstrom, L. *Chem. Rev.* **2002**, *102*, 4501–4523.

(2) Norin, M.; Hult, K.; Mattson, A.; Norin, T. *Biocatal. Biotransform.* **1993**, *7*, 131–147.

(3) Matthews, B. W.; Sigler, P. B.; Henderson, R.; Blow, D. M. *Nature* **1967**, *214*, 652–656.

(4) Quinn, D. M. *Chem. Rev.* **1987**, *87*, 955–979.

(5) Trinh, N. H.; Hoblyn, J.; Mohanty, S.; Yaffe, K. *J. Am. Med. Assoc.* **2003**, *289*, 210–216.

(6) Soreq, H.; Seidman, S. *Nat. Rev. Neurosci.* **2001**, *2*, 294–302.

(7) Schmid, R. D.; Verger, R. *Angew. Chem., Int. Ed.* **1998**, *37*, 1609–1633.

(8) Jaeger, K. E.; Eggert, T. *Curr. Opin. Biotech.* **2002**, *13*, 390–397.

(9) Reetz, M. T. *Curr. Opin. Chem. Biol.* **2002**, *6*, 145–150.

(10) Guo, Z.; Xu, X. *Org. Biomol. Chem.* **2005**, *3*, 2615–2619.

(11) Gupta, R.; Gupta, N.; Rathi, P. *Appl. Microbiol. Biotechnol.* **2004**, *64*, 763–781.

(12) Ban, K.; Kaieda, M.; Matsumoto, T.; Kondo, A.; Fukuda, H. *Biochem. Eng. J.* **2001**, *8*, 39–43.

(13) Botta, B.; Zappia, G.; Tafi, A.; Botta, M.; Manetti, F.; Cernia, E.; Milana, G.; Palocci, C.; Soro, S.; Monache, G. D. *J. Mol. Catal. B: Enzym.* **2002**, *16*, 241–247.

(14) Bornscheuer, U. T.; Kazlauskas, R. J. *Hydrolases in Organic Synthesis: Regio-and Stereoselective Biotransformations*, 2nd ed.; Wiley-VCH: Weinheim, Germany, 2006.

(15) Ghanem, A. *Tetrahedron* **2007**, *63*, 1721–1754.

(16) Gotor-Fernández, V.; Brieva, R.; Gotor, V. *J. Mol. Catal. B: Enzym.* **2006**, *40*, 111–120.

(17) Delinck, D. L.; Margolin, A. L. *Tetrahedron Lett.* **1990**, *31*, 6797–6798.

(18) Stoermer, D.; Caron, S.; Heathcock, C. H. *J. Org. Chem.* **1996**, *61*, 9115–9125.

(19) Schrag, J. D.; Li, Y.; Wu, S.; Cygler, M. *Nature* **1991**, *351*, 761–764.

(20) Kraut, J. *Annu. Rev. Biosc.* **1977**, *46*, 331–358.

(21) Robillard, G.; Shulman, R. G. *J. Mol. Biol.* **1972**, *71*, 507–511.

(22) Robillard, G.; Shulman, R. G. *J. Mol. Biol.* **1974**, *86*, 541–558.

(23) Warshel, A.; Naray-Szabo, G.; Sussman, F.; Hwang, J. K. *Biochem.* **1989**, *28*, 3629–3637.

(24) Ishida, T.; Kato, S. *J. Am. Chem. Soc.* **2004**, *126*, 7111–7118.

(25) Topf, M.; Varnai, P.; Richards, W. G. *J. Am. Chem. Soc.* **2002**, *124*, 14780–14788.

(26) Blow, D. M.; Birktoft, J. J.; Hartley, B. S. *Nature* **1969**, *221*, 337–340.

(27) Hunkapiller, M. W.; Smallcombe, S. H.; Whitaker, D. R.; Richards, J. H. *Biochem.* **1973**, *12*, 4732–4743.

(28) Frey, P. A.; Whitt, S. A.; Tobin, J. B. *Science* **1994**, *264*, 1927–1930.

(29) Fuhrmann, C. N.; Daugherty, M. D.; Agard, D. A. *J. Am. Chem. Soc.* **2006**, *128*, 9086–9102.

(30) Wilmouth, R.; Edman, K.; Neutze, R.; Wright, P.; Clifton, I.; Schneider, T.; Schofield, C.; Hajdu, J. *Nat. Struct. Biol.* **2001**, *8*, 689.

(31) Zhang, Y.; Kua, J.; McCammon, J. A. *J. Am. Chem. Soc.* **2002**, *124*, 10572–10577.

(32) Otte, N. Ph.D. thesis, Universität Düsseldorf, Germany, 2006.

(33) Topf, M.; Richards, W. G. *J. Am. Chem. Soc.* **2004**, *126*, 14631–14641.

(34) Ferraboschi, P.; Casati, S.; De Grandi, S.; Grisenti, P.; Santaniello, E. *Biocatal. Biotransform.* **1994**, *10*, 279–288.

(35) Mezzetti, A.; Keith, C.; Kazlauskas, R. J. *Tetrahedron Asym.* **2003**, *14*, 3917–3924.

(36) Weissfloch, A. N. E.; Kazlauskas, R. J. *J. Org. Chem.* **1995**, *60*, 6959–6969.

(37) Tuomi, W. V.; Kazlauskas, R. J. *J. Org. Chem.* **1999**, *64*, 2638–2647.

(38) Mezzetti, A.; Schrag, J. D.; Cheong, C. S.; Kazlauskas, R. J. *Chem. Biol.* **2005**, *12*, 427–437.

(39) Tomić, S.; Dobovičnik, V.; Šunjić, V.; Kojić-Prodić, V. *Croat. Chim. Act.* **2001**, *74*, 343–357.

(40) Tafi, A.; van Almsick, A.; Corelli, F.; Crusco, M.; Laumen, K. E.; Schneider, M. P.; Botta, M. *J. Org. Chem.* **2000**, *65*, 3659–3665.

(41) Zuegg, J.; Hönig, H.; Schrag, J. D.; Cygler, M. *J. Mol. Catal. B: Enzym.* **1997**, *3*, 83–98.

(42) Ferraboschi, P.; Casati, S.; Manzocchi, A.; Santaniello, E. *Tetrahedron Asymm.* **1995**, *6*, 1521–1524.

(43) Tafi, A.; Manetti, F.; Botta, M.; Casati, S.; Santaniello, E. *Tetrahedron Asymm.* **2004**, *15*, 2345–2350.

(44) Ema, T. *Curr. Org. Chem.* **2004**, *8*, 1009–1025.

(45) Gao, J.; Ma, S.; Major, D. T.; Nam, K.; Pu, J.; Truhlar, D. G. *Chem. Rev.* **2006**, *106*, 3188–3209.

(46) Laio, A.; Parrinello, M. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12562–12566.

(47) Micheletti, C.; Laio, A.; Parrinello, M. *Phys. Rev. Lett.* **2004**, *92*, 170601.

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

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

(50) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

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

(52) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. *J. Comput. Chem.* **1993**, *14*, 1347–1363.

(53) Schrag, J. D.; Li, Y.; Cygler, M.; Lang, D.; Burgdorf, T.; Hecht, H. J.; Schmid, R.; Schomburg, D.; Rydel, T. J.; Oliver, J. D. *Structure* **1997**, *5*, 187–202.

(54) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

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

(56) Grest, G. S.; Kremer, K. *Phys. Rev. A* **1986**, *33*, 3628–3631.

(57) Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. *J. Chem. Phys.* **1995**, *103*, 4613–4621.

(58) Laino, T.; Mohamed, F.; Laio, A.; Parrinello, M. *J. Chem. Theory Comput.* **2005**, *1*, 1176–1184.

(59) Laino, T.; Mohamed, F.; Laio, A.; Parrinello, M. *J. Chem. Theory Comput.* **2006**, *2*, 1370–1378.

(60) Lippert, G.; Hutter, J.; Parrinello, M. *Theor. Chem. Acc.* **1999**, *103*, 124–140.

(61) VandeVondele, J.; Krack, M.; Mohamed, F.; Parrinello, M.; Chassaing, T.; Hutter, J. *Comput. Phys. Commun.* **2005**, *167*, 103–128.

(62) The CP2K developers group. http://cp2k.berlios.de (accessed June 2, 2009).

(63) Ishida, T.; Kato, S. *J. Am. Chem. Soc.* **2003**, *125*, 12035–12048.

(64) Maseras, F.; Morokuma, K. *J. Comput. Chem.* **1995**, *16*, 1170–1179.

(65) Schaefer, A.; Huber, C.; Ahlrichs, R. *J. Chem. Phys.* **1994**, *100*, 5829.

(66) Goedecker, S.; Teter, M.; Hutter, J. *Phys. Rev. B* **1996**, *54*, 1703–1710.

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

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

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

(70) Genovese, L.; Deutsch, T.; Neelov, A.; Goedecker, S.; Beylkin, G. *J. Chem. Phys.* **2006**, *125*, 074105.

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

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

(73) Laio, A.; Rodriguez-Fortea, A.; Gervasio, F. L.; Ceccarelli, M.; Parrinello, M. *J. Phys. Chem. B* **2005**, *109*, 6714–6721.

(74) Klahn, M.; Braun-Sand, S.; Rosta, E.; Warshel, A. *J. Phys. Chem. B* **2005**, *109*, 15645–15650.

(75) Fersht, A. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*; W.H. Freeman and Company: New York, 1999.

(76) Pleiss, J.; Scheib, H.; Schmid, R. D. *Biochimie* **2000**, *82*, 1043–1052.

(77) Gentner, C.; Schmid, R. D.; Pleiss, J. *Colloids Surf., B* **2002**, *26*, 57–66.

(78) Radisky, E. S.; Lee, J. M.; Lu, C. J. K.; Koshland, D. E., Jr. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 6835–6840.

(79) Kim, K. K.; Hwang, K. Y.; Jeon, H. S.; Kim, S.; Sweet, R. M.; Yang, C. H.; Suh, S. W. *J. Mol. Biol.* **1992**, *227*, 1258–1262.

(80) Kim, K. K.; Song, H. K.; Shin, D. H.; Hwang, K. Y.; Suh, S. W. *Structure* **1997**, *5*, 173–185.

(81) Noble, M. E.; Cleasby, A.; Johnson, L. N.; Egmond, M. R.; Frenken, L. G. *FEBS Lett.* **1993**, *331*, 123–8.

(82) Lang, D. A.; Mannesse, M. L. M.; De Haas, G. H.; Verheij, H. M.; Dijkstra, B. W. *Eur. J. Biochem.* **1998**, *254*, 333–340.

(83) Schramm, V. L. *Arch. Biochem. Biophys.* **2005**, *433*, 13–26.

(84) Lafaquiere, V.; Barbe, S.; Puech-Guenot, S.; Guieysse, D.; Cortés, J.; Monsan, P.; Siméon, T.; André, I.; Remaud-Siméon, M. *ChemBioChem* **2009**, *10*, 2760–2771.

CT900636W

JCTC Journal of Chemical Theory and Computation

# A Revised Density Function for Molecular Surface Calculation in Continuum Solvent Models

Xiang Ye, Jun Wang, and Ray Luo*

*Department of Molecular Biology and Biochemistry, University of California, Irvine, California 92697-3900*

Received June 23, 2009

**Abstract:** A revised density function is developed to define the molecular surface for the numerical Poisson−Boltzmann methods to achieve a better convergence and a higher numerical stability. The new density function does not use any predefined functional form but is numerically optimized to reproduce the reaction field energies computed with the solvent excluded surface definition. An exhaustive search in the parameter space is utilized in the optimization, using a wide-range of training molecules including proteins, nucleic acids, and peptides in both folded and unfolded conformations. A cubic-spline function is introduced to guarantee good numerical behavior of the new density function. Our test results show that the average relative energy errors computed with the revised density function are uniformly lower than 1% for both training and test molecules with different sizes and conformations. Our transferability analysis shows that the performance of the new method is mostly size and conformation independent. A detailed analysis further shows that the numerical forces computed with the revised density function converge better with respect to the grid spacing and are numerically more stable in tested peptides.

## Introduction

Solvation is one of the essential determinants of the structure and function of proteins and nucleic acids.[1−14] To model solvation interactions in classical molecular simulations, explicit solvent models that explicitly represent every atom in a solvent molecule are natural choices. However, limitations of explicit solvent models have also been recognized. Apparently, the computational cost is a concern when explicit solvent models are used. Often overlooked are the pre-equilibration and sampling needs. Indeed the exponentially large phase space of explicit solvent degrees of freedom makes the convergence of simulations very challenging. Analyses of these limitations prompted pioneers in molecular simulations to propose implicit representations to model solvation, especially in biomolecular applications.[1−14]

Implicit solvent models replace explicit solvent interactions with an equivalent energetic term based on a mean field approximation. Accuracy and transferability often requires decomposing the mean field potential into electrostatic and nonelectrostatic components and modeling the two components separately. Such an approach can reduce the computation cost but has been found to provide a certain degree of accuracy in the treatment of solvation interactions.[1−14] With over 20 years of developments, implicit solvent models, especially those based on the Poisson−Boltzmann (PB) theory, have been widely accepted in studies of solvation interactions. In the PB-based implicit solvent models, the electrostatic interaction or, more fundamentally, the electrostatic potential is assumed to obey the classical PB equation:

$$\nabla \cdot [\varepsilon \nabla \phi] = -\rho - \lambda \sum q_i n_i^0 \exp[-\beta q_i \phi] \qquad (1)$$

where $\varepsilon$ is the dielectric constant, $\phi$ is the electrostatic potential, $\rho$ is the solute charge density, $\lambda$ has a value of 0 wherever mobile ions cannot penetrate and a value of 1 where they can, $n_i^0$ is the number density of mobile ions of type $i$ in the bulk solution, $q_i$ is the charge of the mobile ions of type $i$, and $\beta = 1/kT$. Here $k$ is the Boltzmann constant, and $T$ is the temperature. When the Boltzmann factor is close to zero, eq 1 can be linearized as

* Corresponding author. Telephone: (949) 824-9528. E-mail: ray.luo@uci.edu.

10.1021/ct900318u © 2010 American Chemical Society
Published on Web 03/01/2010

$$\nabla \cdot [\varepsilon \nabla \phi] = -\rho + \lambda \sum \beta q_i^2 \phi n_i^0 \qquad (2)$$

Adaptation of the PB solvent models to molecular simulations requires a numerical solution of the three-dimensional (3-D) partial differential equation. However, the numerical procedure has been a bottleneck, largely limiting the application to calculations with static structures only. The difficulty lies in the numerical procedure to apply such solvent models, which involves discretization of the partial differential equation into a system of linear or nonlinear equations that tends to be rather large; it is not uncommon to have millions of unknowns in biochemical applications. In addition, the setup of the system before the numerical solution and postprocessing to obtain energies and forces are both nontrivial. Three major discretization methods are widely used in biomolecular applications. The most commonly used approach is the finite-difference method.[15−32] In this method, the physical properties of the solution, such as atomic charges and dielectric constants, are mapped onto rectangular grid points, and a discrete approximation to the governing partial differential equation is produced. The second approach is the finite-element method,[33−38] which approximates the potential by a superposition of a set of basis functions. A linear or nonlinear system for the coefficients produced by the weak formulation has to be solved. The third approach is the boundary-element method.[39−52] In the boundary-element method, the Poisson or PB equation is solved for either the induced surface charge[39−41,43,45,46,49,52] or the normal component of the electric displacement[42,44,47,48,50,51] on the dielectric boundary between the solute and the solvent.

Due to the computational expense for solving the PB equation numerically, considerable efforts have been invested in approximating the solution of the PB equation via methods, such as the semianalytical generalized Born (GB) model,[53−63] the induced multipole model,[64] the dielectric screening model,[65,66] and others. The pairwise GB model, in particular, has been widely accepted as an efficient estimation of the solution of the PB equation, as recently reviewed.[5,6,9−11]

A crucial component of all implicit solvent models within the PB framework is the dielectric model, i.e., the dielectric constant distribution of a given solution system. Typically, a solution system is divided into the low-dielectric interior and the high-dielectric exterior by a molecular surface. That is to say that the molecular surface is used as the dielectric interface between the two piece-wise dielectric constants. In numerical PB calculations, such as the finite-difference methods, discretization of the molecular surface is required. One possible approach is to build the molecular surface analytically and then to map it onto a grid.[67−69] However, analytical procedures can be quite time-consuming and do not necessarily offer any advantages for finite-difference calculations because the surface must, in any case, be mapped onto a grid lattice. In this study, we focus on representations of molecular surface for numerical solutions of the PB equation.

In numerical solutions of the PB equation, the solvent excluded surface (SES) is the most used surface definition.[24,26] Indeed, recent comparative analyses of both PB-based and TIP3P solvent models show that the SES definition is reasonable in calculation of reaction field energies and electrostatic potentials of mean force of hydrogen-bonded and salt-bridged dimers with respect to the TIP3P explicit solvent.[70−72] However, a previous test of the SES definition in the finite-difference solution indicates that it is numerically unstable for molecular dynamics.[73] Similar numerical difficulty was also observed in the pairwise GB method when the SES definition was used.[74] A major limitation of the SES definition is the re-entry volume; it is found that in simulations of proteins at room temperature, large re-entry volumes generated by nonbonded atoms come and go as often as every femtosecond when the nearby atoms undergo vibrational motion.[73] Thus, extremely large surface derivatives with respect to atomic coordinates may occur in the SES definition. In addition, surface cusp may also exist given certain combinations of atom and probe radii and arrangements of atoms.

The van der Waals (VDW) surface, or the hard sphere surface, represents the low-dielectric molecular interior as a union of atomic VDW spheres. With the VDW definition, surface derivatives with respect to the atomic coordinates are much better behaved, different from the SES definition. However, surface cusp at the joint between any two spheres may still cause instability in numerical solutions and in force interpretations, just like the situation in the SES definition. In addition, there exist many nonphysical high (solvent) dielectric pockets inside the solute interior when the VDW definition is used, as discussed in ref 75. These small buried "solvent pockets" result in a molecular interior too hydrophilic, which would cause proteins to unfold. In addition, the complex dielectric interface due to the buried solvent pockets also results in an unsmooth field distribution, leading to unstable dynamics simulations. Considering these limitations, the modified VDW definition was proposed. The basic idea of the modified VDW definition is to use the solvent accessible surface (SAS) definition for fully buried atoms and the VDW definition for fully exposed atoms.[73] This is realized with a set of conformation-dependent modified VDW radii, whose calculation requires the solvent accessible surface area of all atoms to determine their solvent accessibility.[73] Apparently the definition of modified VDW radii has to be smooth to be any use for dynamics simulations. The standard VDW surface can then be generated with the modified VDW radii. The harmonic dielectric smoothing is also applied to smooth the dielectric transition between solvent and solute.[76] Apparently, the dielectric distribution within and around buried atoms is very smooth, i.e., it is all part of the solute low dielectric. However, the dielectric distribution around exposed atoms can still show spatial fluctuation, as in the original VDW surface. Thus, an additional step in the modified VDW definition is used to smooth the spatial fluctuation around exposed atoms,[73] though it is difficult to implement and hard to be optimized to reproduce the SES.

The density approaches have recently been developed and can be used for numerical PB solutions. Either a Gaussian-like function or a smoothed step function has been explored in previous developments.[77,78] In these approaches, a distance-dependent density/volume exclusion function is used to define each atomic volume or dielectric constant directly. This is in contrast to the hard-sphere definition of atomic volume, as in the VDW or the SES definition. Note that the use of a smooth function allows the "boundary region" or the "solute/solvent transition region" extends both inward and outward, i.e., the abrupt dielectric transition in the classical two-dielectric model has been replaced with a smooth dielectric transition region of finite width. Indeed, extension inward is necessary in the new definition to reproduce the results of the classical two-dielectric energetics for small training molecules.[77,78] In doing so, the trailing tail outside the cavity radii can be used to smooth out the small cracks and crevices formed by neighboring atoms. Apparently, any surface cusps are removed by the use of smooth density functions. However, their agreements with the classical two-dielectric model for typical macromolecules are not very good, as will be shown below even if the agreements can be excellent for the small training molecules.[77,78]

In this development, we have revisited the density function strategy by combining it with the modified VDW definition to balance numerical stability and model quality for numerical PB applications. Specifically, we have optimized the new density function to reproduce, as much as possible, the reaction field energies based on the SES definition due to its reasonable agreement with explicit solvent models. A large and diversified training set of biomolecules is used during the optimization. A separate test set of biomolecules is also used to validate its performance. Its benefit in improving the convergence and the numerical stability of electrostatic force calculations is also discussed.

## Methods

**Finite-Difference Method.** In this work, we focus on numerical solution of the linearized PB equation. However, the proposed new numerical surface procedure can certainly be applied to the full nonlinear PB equation, which has been implemented in the Amber/PBSA program.[25,30−32] A widely used numerical method is the finite-difference method. It requires mapping the problem domain onto a lattice of grid points. The grid points are connected with grid edges. Solute atomic charge distribution is mapped onto grid points, and the dielectric constant distribution is mapped onto the centers of grid edges. The linearized PB equation can then be converted into a linear system with the finite-volume scheme. Under this discretization scheme, the partial differential equation can be written as follows at each grid point:

$$\begin{aligned}
&-h^2\varepsilon\left(i-\frac{1}{2},j,k\right) && [\phi(i-1,j,k)-\phi(i,j,k)] \\
&-h^2\varepsilon\left(i+\frac{1}{2},j,k\right) && [\phi(i+1,j,k)-\phi(i,j,k)] \\
&-h^2\varepsilon\left(i,j-\frac{1}{2},k\right) && [\phi(i,j-1,k)-\phi(i,j,k)] \\
&-h^2\varepsilon\left(i,j+\frac{1}{2},k\right) && [\phi(i,j+1,k)-\phi(i,j,k)] \\
&-h^2\varepsilon\left(i,j,k-\frac{1}{2}\right) && [\phi(i,j,k-1)-\phi(i,j,k)] \\
&-h^2\varepsilon\left(i,j,k+\frac{1}{2}\right) && [\phi(i,j,k+1)-\phi(i,j,k)] \\
&+\kappa^2 && \phi(i,j,k) && = h^{-3}q(i,j,k)
\end{aligned} \tag{3}$$

where $h$ is the spacing in each dimension, $i$, $j$, and $k$ are the grid indexes along the $x$, $y$, and $z$ axes, respectively, and $\varepsilon(i \mp {}^1/_2, j, k)$ is the dielectric constant between grids $(i, j, k)$ and $(i \mp 1, j, k)$. Both $\varepsilon(i, j \mp {}^1/_2, k)$ and $\varepsilon(i, j, k \mp {}^1/_2)$ are defined similarly. In the Boltzmann term, $\kappa^2$ absorbs all the related coefficients, and $q(i, j, k)$ is the total charge within the cubic volume centered at $(i, j, k)$. We can use several methods to solve the linear system, such as Gauss−Seidel, Jacobi, successive over relaxation, conjugate gradient, and so on.[79,80]

**Dielectric Distribution Model.** A key issue in the solution of eq 3 is how to map the dielectric constant distribution on all grid edges. In biomolecular calculations, the dielectric constant distribution often adopts a piece-wise constant model, where the dielectric within the molecular surface is assigned to that of the solute and the dielectric outside the molecular surface is assigned to that of the solvent. Within this model, the dielectric constant on a grid edge, apparently, should be assigned to the dielectric constant in this region where the two neighbor grid points belong. However, when the two neighbor grid points belong to different dielectric regions, i.e., when the grid edge is a boundary grid edge, its dielectric constant is nontrivial to assign because the dielectric constant is discontinuous across the interface. One simple treatment is the use of harmonic average (HA) of the two dielectric constants at the center of grid edges across the solute/solvent boundary.[76] For example, if $(i - 1, j, k)$ and $(i, j, k)$ belong to solute and solvent regions, respectively, then there must be an intersection point on the grid edge between $(i - 1, j, k)$ and $(i, j, k)$. Denote $a$ as the distance from the intersection point to the grid point $(i - 1, j, k)$ and $b$ as the distance from the same intersection point to the grid point $(i, j, k)$. In HA, $\varepsilon(i - 1/2, j, k)$ is defined as

$$\varepsilon\left(i-\frac{1}{2},j,k\right)=\frac{h}{\dfrac{a}{\varepsilon(i-1,j,k)}+\dfrac{b}{\varepsilon(i,j,k)}} \tag{4}$$

This strategy has been shown to improve the convergence of reaction field energies with respect to the grid spacing and to reduce the grid dependence of the solvation energetics.[76] The smoothed dielectric constant transition across the solute/solvent interface also makes it possible to compute dielectric boundary force via a variational approach proposed by Gilson et al.[81] Apparently, only the intersection points between the molecular surface and the boundary edges are needed to utilize eq 4 to assign dielectric constants at boundary grid edges.

**A Revised Density Function Strategy.** In this study, we explored representing the molecular surface *indirectly* or *implicitly* with a molecular density function. That is to say that the two-dimensional (2-D) molecular surface is represented as an equi-density surface of a 3-D density function. The strategy appears to complicate the numerical problem by increasing the dimensionality of the procedure. However, the primary aim is to reduce the numerical instability arising from using hard spheres in classical molecular surface representation. In addition, the density function strategy naturally fits in the level set method that can be utilized to compute, numerically, many properties of the molecular surface in a rather straightforward manner for the finite-difference methods, as will be described.

The central idea of our revised density function is the same as previous attempts to apply density functions to molecular surface or volume presentations,[77,78] i.e., to use a function that smoothly maps out the interior of the SES of a molecule. However, we require here that it also reproduces the solvation energetics computed with the hard-sphere-based SES definition as much as possible. Specifically, atoms are described by atomic density functions, and these are combined using a composite molecular density function that can be used to calculate discretized molecular surface, i.e., the intersection points between the molecular surface and the boundary grid edges. As discussed above, these points are used to define the dielectric distribution of the solution system by the harmonic average method.[76]

Specifically, given the $n^{\text{th}}$ atom centered at $\mathbf{r}_n$, its density $\rho_n$ is defined as

$$\rho_n = \rho_n(x) \tag{5}$$

with $x = (d - r_c)/2r_p$. Here, $d = |\mathbf{r} - \mathbf{r}_n|$ is the distance to the atomic center ($\mathbf{r}_n$), $r_c$ is the VDW radius of the atom, and $r_p$ is the solvent probe radius. The domain of the independent variable, $x$, is then set to be $[-r_c/2r_p, 1]$, where $x = -r_c/2r_p$ corresponds to the atomic center and $x = 1$ corresponds to one probe diameter away from the atomic VDW surface. Note also that $x = 0$ corresponds to the atomic VDW surface, when $x$ is so defined. Apparently the domain of the dependent variable, the density value ($\rho_n$), needs to satisfy certain constraints to be physical or reasonable and to be smooth for numerical PB methods. Here, the density function is required to satisfy:

$$\begin{aligned} &\rho_n(x) > 1, \quad x < 0 \\ &\rho_n(x) = \begin{cases} 1, & x = 0 \\ 0, & x = 1 \end{cases} \end{aligned} \tag{6}$$

The constraints are meant to specify that: (i) the density function is always positive; and (ii) the density value within any atomic VDW surface is guaranteed to be ≥1 within a molecule, as will become clear below. A main point of this study is not to use any prescribed function form but to optimize a numerical function, i.e., a table lookup function, which satisfies the above constraints, and to achieve the best possible agreement with a given benchmark for a specific training set. The details of the benchmark, the training set, and the quality measure will be discussed below. Of course, to guarantee smoothness and good numerical behaviors the cubic-spline interpolation is used to interpolate the function value within the allowed range of $x$.

With the definition of atomic density functions, we can now define a molecular density function as

$$\rho_{\text{mol}}(\mathbf{r}) = 1 - \prod_n (1 - \rho_n) \tag{7}$$

in terms of a repeated product over atomic density functions.[77,78] Expanding this expression we obtain

$$\begin{aligned} \rho_{\text{mol}}(\mathbf{r}) &= \sum_n \rho_n - \sum_{n>m} \rho_n\rho_m + \sum_{n>m>l} \rho_n\rho_m\rho_l + ... \\ &= \rho_{\text{sum}}(\mathbf{r}) + \text{''intersection terms''} \end{aligned} \tag{8}$$

where $\rho_{\text{sum}}(\mathbf{r}) = \sum_n \rho_n$ is the linear summation of all atomic terms, and the higher products, i.e., the "intersection terms", represent corrections for over or under counting of atomic intersections.[77,78] Use of $\rho_{\text{mol}}(\mathbf{r})$ would produce a molecular volume almost the same as the "classical" VDW volume, which is not desired in numerical PB applications as reviewed in the introduction.[77,78] The original density function ($\rho_{\text{mol}}$) was defined in such a way that $\rho_{\text{mol}} > 1$ corresponds the VDW volume. If only the leading term $\rho_{\text{sum}}$ were used, then $\rho_{\text{sum}} > 1$ would correspond to an overestimation of the VDW volume. Corrections from the intersection terms in $\rho_{\text{mol}}$ were used to successively correct the volume so that it eventually leads to a volume consistent with the VDW volume. That is to say that when only the leading term is used, the definition of $\rho_{\text{sum}} > 1$ has the potential to capture the "re-entry volume", in addition to the VDW volume when properly optimized. A similar idea was used in previous work.[77,78]

**Combination with the Modified VDW Surface.** Note that the molecular volume defined by a density function crucially depends upon the function form. A function defined to decay faster to zero would lead to a smaller molecular volume, i.e., more solvent-exposed interatomic crevices would exist. In contrast, a function defined to decay slower to zero would lead to a larger molecular volume, i.e., less solvent-exposed interatomic crevices would exist. Our initial analysis of the density function approach shows that its sensitivity to the function form makes it very difficult to reproduce the SES definition.

The limitation can be attributed to the different requirements between defining the solvent-exposed surface for exposed atoms and the solvent-excluded volume for buried atoms. Our analysis shows that the density function has to be defined to decay to zero faster to capture the solvent-exposed surface in the SES definition. However, the density function has to be defined to decay to zero slower to capture the solvent-excluded volume in the SES definition. Since more atoms on peptides are solvent-exposed, while more atoms in proteins are solvent-excluded, the density function tends to make the proteins too hydrophilic, if it is optimized with respect to the peptides. In contrast, the density function tends to make the peptides too hydrophobic, if it is optimized with respect to the proteins. Overall, it is too difficult to find a compromise that would work for different sized molecules.

We explored combining the density function with the modified VDW approach reviewed in the introduction to overcome its limitation. Specifically, the density function is revised as follows:

A. Compute SAS with unmodified VDW radii ($r_c$).

B. Compute modified VDW radii ($r_c^{\mathrm{m}}$) based on atomic SAS (see below).

C. Calculate the composite density function, $\rho_{\mathrm{sum}}(\mathbf{r}) = \sum_n \rho_n$, with the modified VDW radii, i.e., with $r_c$ replaced by $r_c^{\mathrm{m}}$ in eq 5.

The modified VDW radius of atom $i$ is defined to be smoothly dependent upon atomic SAS as follows:

$$r_{c,i}^{\mathrm{m}} = \begin{cases} r_{c,i} + \frac{1}{2}\eta r_{\mathrm{p}}\left[1 + \cos\left(\pi \frac{A_i^{\mathrm{s}}}{A^{\mathrm{c}}}\right)\right], & A_i^{\mathrm{s}} < A^{\mathrm{c}} \\ r_{c,i}, & A_i^{s} \geq A^{c} \end{cases} \qquad (9)$$

where $r_{c,i}^{\mathrm{m}}$ is the modified radius of atom $i$, $r_{c,i}$ is the unmodified VDW radius of atom $i$, $\eta \in [0,1]$ controls how much increment to be added to the unmodified VDW radius, $r_{\mathrm{p}}$ is the solvent probe radius, $A_i^{\mathrm{s}}$ is the relative solvent accessibility of atom $i$, and $A^{\mathrm{c}} \in [0,1]$ is the cutoff relative solvent accessibility. Equation 9 shows that only VDW radii of buried or somewhat exposed atoms are incremented, while VDW radii of highly exposed atoms are not incremented.[73]

Thus, when the density function is defined with the modified VDW radii, we can achieve the goal of raising the density function value higher within the molecular interior by effectively pushing the atomic VDW surface outward, while still preserving the low-density values for solvent-exposed atoms that are needed to define the solvent-exposed surface. As will be shown below, a more consistent performance can be realized among different sized molecules when the modified VDW radii are used in the density function approach. Of course, parameters $\eta$ and $A^{\mathrm{c}}$ should be optimized, along with the density function, to best reproduce the reaction field energies with the SES definition.

**Use of Numerical Function to Represent the Atomic Density Function.** Up to this point, we have yet to define the atomic density function, except its overall properties. As discussed, a main point of this study is not to use any prescribed function form but to optimize a numerical function for $\rho_n(x)$ to achieve the best possible agreement between $\rho_{\mathrm{sum}}(\mathbf{r})$ and the given benchmark for a specific training set. Specifically, five intervals in the range of [0, 1] were used to optimize the numerical function. These are then interpolated over the domain of interest with the cubic-spline interpolation to guarantee a continuous and smooth $\rho_n(x)$.[82] It is found that use of more intervals does not improve the quality of the agreement with the SES definition. Of course, the use of fewer intervals reduces the quality of the agreement.

Apparently, it is perfectly reasonable to include intervals <0 in the optimization, but these intervals are already within the atomic VDW volumes, which are guaranteed to be within the molecular surface and do not contribute to the optimization quality of the function. Inclusion of these intervals only increases the difficulty of the optimization problem. Instead, we extended the density function all the way to the atomic center with a linear function with the slope of the cubic-spline function at $x = 0$. For example, the atomic density function optimized within the range of [0, 1] is shown in Figure 1, which also plots two more intervals <0. Note also that the numerical function within the first interval [0, 0.2] is almost linear, even if the cubic-spline interpolation is used (Figure 1).

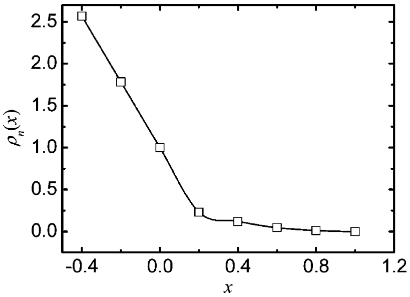


***Figure 1.*** Optimized atomic density function.

As described above, the numerical density function is optimized with a benchmark set. Here the PB reaction field energies in the SES definition were used in this study. Our choice of the SES definition as a benchmark was based on recent comparative analyses of both PB and TIP3P solvent models. These studies show that the SES definition is a reasonable surface definition in the calculation of reaction field energies and electrostatic potentials of mean force of hydrogen-bonded and salt-bridged dimers, with respect to the TIP3P explicit solvent.[70−72] Finally, to ensure transferability and universality of the density function, we used a large set of diversified biomolecules, both folded and unfolded conformations, in the optimization of the density function.

In summary, the following parameters were optimized: (i) the four density function values at $x = 0.20$, 0.40, 0.60, and 0.80, respectively, to determine $\rho_n(x)$; and (ii) $\eta$ and $A^{\mathrm{c}}$ to determine the optimal modified VDW radii (eq 9). The optimization was conducted with respect to the benchmark reaction field energies computed with the SES definition. The average error (i.e., unsigned relative deviation) between the two sets of reaction field energies was used as the optimization measure. Note that the function values at $x = 0$ and 1 have been fixed as 1.0 and 0.0, respectively, according to eq 6. In this study, we used a three-step systematic scan of the parameter space to optimize the parameters. At step one, an initial scan in the resolution of 0.2 was used for all six parameters. At step two, a refinement scan in the resolution of 0.05 was used in the reduced search space (±0.2), centered on the best parameter set from step one. Finally, at step three, a second refinement scan in the resolution of 0.01 was used in the reduced search space of (±0.05), centered on the best parameter set from step two. The optimized parameters are shown in the Results and Discussion Section.

**Implicit Molecular Surface Representation by the Level-Set Method.** Once $\rho_{\mathrm{sum}}(\mathbf{r})$ is defined on grid points, we need to know where the equi-density surface, $\rho_{\mathrm{sum}}(\mathbf{r}) = 1$, intersects all grid edges to use the harmonic average method to set up the dielectric constants of all boundary grid edges, see eq 4. In addition, we also need the surface normal direction in the force calculation, as will be shown next.

We explored computing these numerical surface properties with the level set method.[82,83] In the level set method, a scalar function, $\varphi(\mathbf{r})$, i.e., the level set function, is used to represent the surface implicitly. For example, to describe a 2-D spherical surface with the radius of 1, we can define a level set function as $\varphi(\mathbf{r}) = |\mathbf{r}|^2 - 1$ in the 3-D space. Thus, the spherical surface consists of all the points satisfying the condition of $\varphi(\mathbf{r}) = 0$. In addition, for any point $\mathbf{r}$, we can easily determine it is within the surface, if $\varphi(\mathbf{r}) < 0$ or outside of the surface, if $\varphi(\mathbf{r}) > 0$. It should be pointed out that for a specific surface ($\varphi(\mathbf{r}) = 0$), there may be many different level set functions because the requirement of $\varphi(\mathbf{r}) = 0$ cannot uniquely define the 3-D function. To utilize the level set method to define the molecular surface, we need to define a level set function so that $\varphi(\mathbf{r}) = 0$, i.e., the zero level set corresponds to the molecular surface. In addition, positive and negative function values denote the solvent and the solute sides of the surface, respectively. Thus, our revised density function can be used to define a level set function as

$$\varphi(\mathbf{r}) = 1 - \rho_{\mathrm{sum}}(\mathbf{r}) \tag{10}$$

With these preparations, we proceed to compute the interaction point of a boundary grid edge and a molecular surface, as follows. Without loss of generality, suppose that this is an $x$-edge flanked by two grid points $x_1$ and $x_2$. The level set function values are $\varphi_1$ and $\varphi_2$, respectively. Apparently, we have $\varphi_1 \times \varphi_2 < 0$, since the sign of the level set function, defined by eq 10, changes when crossing the molecular surface and the intersection point is between $x_1$ and $x_2$. Next, we choose a third grid point, $x_3$, from the two grid points that flank the $x$-edge in the $x$-direction, i.e., it is right next to $x_1$ or $x_2$ and has the same $y$ and $z$ coordinates. Given the three grid points with their $x$ coordinates as $x_1$, $x_2$, $x_3$ and with corresponding level set functions as $\varphi_1$, $\varphi_2$, $\varphi_3$, respectively, a quadratic function $\varphi = a_2x^2 + a_1x + a_0$ can be determined to pass through three points ($x_1$, $\varphi_1$), ($x_2$, $\varphi_2$), and ($x_3$, $\varphi_3$). Thus, the intersection point is simply the root of the quadratic equation $a_2x^2 + a_1x + a_0 = 0$, within [$x_1$, $x_2$]. It has been shown that the error in the calculated intersection point scales as $O(h^2)$.[82,83]

With the level set method, calculation of other surface properties is also straightforward. In this study, the surface normal direction is needed at every intersection point and can be computed as[82,83]

$$\mathbf{n} = \frac{\nabla\varphi}{|\nabla\varphi|} = \frac{(\varphi_x, \varphi_y, \varphi_z)}{(\varphi_x^2 + \varphi_y^2 + \varphi_z^2)^{1/2}} \tag{11}$$

where $\varphi_x$ is the derivative of $\varphi$, with respect to $x$. Other symbols are defined similarly. The simplicity of the level set method makes it well suited to the finite-difference method, where these derivatives can be interpolated with accuracy of $O(h^2)$.[82,83]

## Computation Details

**Electrostatic Energy and Force Calculation.** The electrostatic reaction field energy was computed via the dielectric polarization charges, which were calculated using the Gauss law and the grid potential obtained from the finite-difference solution of the PB equation.[26] In the finite-difference method, the dielectric polarization charges are located on the boundary grid points, i.e., the grid points surrounded by nonuniform dielectric grid edges. To improve the convergence of reaction field energy, the polarization charges were first projected onto the molecular surface, according to the procedure described by Rocchia et al.,[26] before they were used to compute the reaction field energy as a pairwise summation of Coulombic interactions between atomic and polarization charges.

It is well-known that the electrostatic force density can be derived through the divergence of the Maxwell stress tensor ($\mathbf{P}$) as[84,85]

$$\begin{aligned}\mathbf{f} = \nabla\cdot\boldsymbol{P} &= \frac{\partial}{\partial x}(i\cdot\boldsymbol{P}) + \frac{\partial}{\partial y}(j\cdot\boldsymbol{P}) + \frac{\partial}{\partial z}(k\cdot\boldsymbol{P}) \\ &= \rho^f\mathbf{E} - \frac{1}{8\pi}E^2\nabla\varepsilon - \Delta\Pi\nabla\lambda\end{aligned} \tag{12}$$

where $\rho^f$ is the fixed charge density, $\mathbf{E}$ is the electric field, $\varepsilon$ is the dielectric constant, $\Delta\Pi$ is the excess osmotic pressure,[86] and $\lambda$ is the Stern layer defined so that it is 1 in regions accessible to the mobile ions and 0 elsewhere. This is consistent with the formulation of Gilson et al.[81]

Equation 12 shows that there are three components in the total electrostatic forces: (i) the Coulombic and reaction field forces acting on the atomic charges, $\rho^f\mathbf{E}$; (ii) the dielectric boundary forces, or pressure acting on the dielectric boundary, $-(1/8\pi)E^2\nabla\varepsilon$; and (iii) the ionic boundary forces, or pressure on the ionic boundary. Since the Coulombic forces can be computed analytically by pairwise summation of Coulombic interactions among atomic charges, only the rest of the force components were computed numerically.

Similar to the treatment of reaction field energy, dielectric polarization charges can be used to improve the convergence of reaction field forces, with respect to the grid spacing. Specifically, the reaction field forces were calculated by the pairwise summation of the Coulombic interactions between polarizarion and atomic charges.[26] The computation of dielectric boundary forces requires the derivative of the dielectric constant. Thus, only a smoothly varied dielectric model, such as eq 4, can be used in eq 12,[85] where the dielectric constants on grid edges across the dielectric boundary (molecular surface) were assigned as weighted harmonic averages of solvent and solute dielectric constants. The finite-difference procedure to implement $-(1/8\pi)E^2\nabla\varepsilon$, given the weighted harmonic averages of dielectric boundary grid edges, has been described in detail by Gilson et al.[81] Note that the dielectric boundary force element is always along the direction of the gradient of dielectric constant, which is the normal direction of the molecular surface. Thus, computation of the dielectric boundary force element requires the numerical calculation of surface normal vectors, which was computed by eq 11.

Due to the fact that the SES surface is not differentiable, the dielectric boundary force elements are distributed to nearby atoms in an ad hoc manner, as follows. For the contact portion of the SES, the surface force elements are distributed to the closest atom sphere. For the re-entry portion of the SES, the dielectric boundary force elements are distributed to the two nearest atom spheres proportional to the inverse

***Table 1.*** Training (1−3) and Test Sets (4−8) Used in This Study[a]

| structure set | no. structures | no. residues | rmsd (Å) |
|---|---|---|---|
| 1: PDB1 | 290 | 16−517 | 0.00 |
| 2: PGB | 314 | 56 | 0.01−13.80 |
| 3: HPN1 | 300 | 16 | 0.02−11.63 |
| 4: PDB2 | 289 | 19−497 | 0.00 |
| 5: ENH | 261 | 54 | 0.02−10.50 |
| 6: SHG | 284 | 57 | 0.16−13.22 |
| 7: helix | 501 | 19 | 0.02−8 0.95 |
| 8: HPN2 | 451 | 16 | 0.02−11.28 |

[a] Root-mean-square deviation from the experimental native structure is defined as rmsd.

of the distances of the two atom spheres. This procedure is also used to distribute the dielectric boundary force elements for the surface defined by the revised density function.

Finally, the ionic boundary forces are $\sim O(10^{-2})$ smaller than those of the reaction field forces and the dielectric boundary forces in water,[81] so that we only focus on the reaction field forces and dielectric boundary forces in the performance analysis of the revised density function below. In addition their performance is apparently more closely related to how the Stern layer is defined, which is beyond the scope of this development.

**Training and Test Sets.** To cover the highly heterogeneous molecular surface topologies, eight different structure sets were used in this study to calibrate and to test the revised density function (see Table 1). Sets 1 and 4 contain a large set of Protein Data Bank (PDB) structures, ranging from small peptides to very large biomolecules (with more than 550 residues) and covering a wide variety of native protein folds.[87] Sets 2, 5, and 6 are unfolded protein conformations of three small global proteins from high-temperature unfolding simulations: 1PGB (alpha/beta), 1ENH (all-alpha), and 1SHG (all-beta). Sets 3, 7, and 8 are from unfolding simulations of three peptides. Set 3 is a hairpin from 1PGB, set 7 is a helix from 1PGB, and set 8 is a hairpin from 1SHG. These sets are used to test how well reaction field energies are reproduced for different native and nonnative conformations of the same protein. Finally, sets 1−3 were used as training sets, and sets 4−8 were used as test sets. All molecular structures were processed with Leap in Amber9[88] and held static in all calculations. As described above, the optimization of the density function was conducted with the reaction field energies of chosen training molecules. In the surface potential analysis, the native structures of 1PGB, 1ENH, and 1SHG and the p53 DNA binding domain with and without DNA were used.

**Other Details.** In all calculations, the ion concentration was set to 0. The dielectric constant of the solvent was set to 80, i.e., for water, while the solute was set to 1. The solvent probe was set to be 0.6 Å. Our use of an unusually small probe radius was based on our previous comparative analyses of the numerical PB and TIP3P solvent models. In these analyses, the reaction field energies of small molecules were found to be not very sensitive to the different probe sizes, but the electrostatic potentials of mean force of the hydrogen-bonded or the salt-bridged dimers were quite sensitive to the probe radius used, and a solvent probe radius of 0.6 Å was found best to reproduce the TIP3P solvent among the tested values.[72] Subsequent analysis of the ion pairs on peptides and proteins also indicates that the probe radius of 0.6 Å can best reproduce the TIP3P solvent [Tan and Luo, manuscript in preparation]. The finite-difference grid spacing was set to be 0.5 Å, if not mentioned otherwise. A two-level electrostatic focusing was used to speed up the assignment of electrostatic boundary condition. The coarse grid spacing was 2.0 Å. The dimension of the coarse grid was set to be twice as large as the dimension of the solute to secure the quality of the nonperiodic boundary condition in the PB calculations. The finite-difference convergence criterion was set to be $10^{-3}$. All other parameters are set to be default as in the PBSA program of Amber 9.[25,73,88]

***Table 2.*** Average Error, Maximum Error, And Error Spread (Average/Max/Spread) for Each of the Training and Test Sets[a]

| structure set | revised density | density | MVDW |
|---|---|---|---|
| 1: PDB1 | 0.47/1.82/3.11% | 1.26/6.58/9.40% | 2.01/6.66/8.07% |
| 2: PGB | 0.76/2.38/3.05% | 0.57/3.50/4.77% | 4.39/6.75/5.37% |
| 3: HPN1 | 0.76/3.26/5.90% | 1.84/3.72/5.36% | 4.37/9.32/10.47% |
| 4: PDB2 | 0.45/2.40/3.86% | 1.27/8.44/10.85% | 1.95/5.95/6.83% |
| 5: ENH | 0.58/1.72/2.50% | 0.56/2.23/2.90% | 3.18/5.27/4.39% |
| 6: SHG | 0.89/2.47/3.01% | 0.71/2.98/4.32% | 4.72/7.42/6.11% |
| 7: helix | 0.54/1.87/3.66% | 1.29/2.82/5.61% | 3.88/7.67/7.46/% |
| 8: HPN2 | 0.58/3.38/5.53% | 1.24/3.12/5.95% | 3.89/8.52/9.22% |

[a] Revised density: density function definition with modified VDW radii. Density: density function definition with unmodified VDW radii. MVDW: VDW surface definition with modified VDW radii.

## Results and Discussion

**Optimization of the Revised Density Function.** As described in the method, the revised density function was optimized by a systematic search of the six parameters with respect to a benchmark set of reaction field energies computed with the SES definition. The optimized parameters, with the lowest average unsigned error with respect to the benchmark set, are as follows: $\rho_n(x) = 0.21$, 0.15, 0.05, and 0.01 at $x = 0.2$, 0.4, 0.6, and 0.8, respectively; $\eta = 0.57$; and $A^c = 0.08$. Once the discrete density function values are given, the cubic-spline function is utilized to interpolate the density function within the range of [0, 1] for $x$. Note that $x = 0$ corresponds to the atomic VDW surface, and the atomic density function is defined for $x < 0$ as a linear function with the slope at $x = 0$. Figure 1 plots the cubic-spline interpolated function with the optimized discrete density function values. To fully understand the performance gain in combining the density function with the modified VDW surface, the pure density function approach, i.e., with density function defined with the unmodified VDW radii, was also optimized with respect to the same benchmark set. The optimized parameters are as follows: $\rho_n(x) = 0.26$, 0.21, 0.05, and 0.01 at $x = 0.2$, 0.4, 0.6, and 0.8, respectively.

It can be seen in Table 2, the revised density function with the modified VDW radii, denoted as "revised density", can achieve average errors less than 1% (PDB1: 0.47; PGB: 0.76; and HPN1: 0.76%) for all three training sets, a respectable agreement. The corresponding values of the density function with the unmodified VDW radii, denoted as "density",

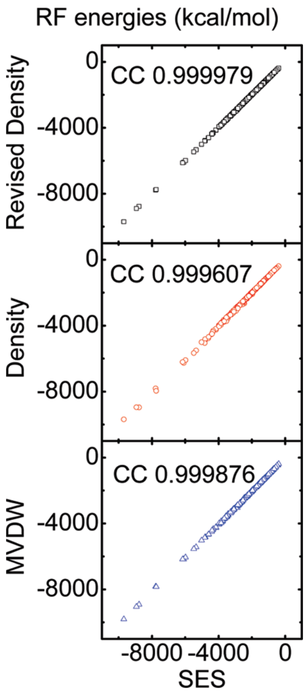


*Figure 2.* Top: Correlation between reaction field (RF) energies computed with the SES definition and those computed with the revised density definition (correlation: 0.999979, slope: 0.999505, offset: 5.30848, and rms relative deviation: 0.57%). Middle: Correlation between RF energies computed with the SES definition and those computed with the density definition (correlation: 0.999607, slope: 1.01631, offset: 11.1488, and rms relative deviation: 1.79%). Bottom: Correlation between RF energies computed with the SES definition and those with the MVDW definition (correlation: 0.999876, slope: 1.01112, offset: −15.306, and rms relative deviation: 2.17%).

(PDB1: 1.26; PGB: 0.57; and HPN1: 1.84%) and modified VDW definition, denoted as "MVDW", are relatively larger (PDB1: 2.01; PGB: 4.39; and HPN1: 4.37%). The average errors of the revised density definition are also lower than those of the density and the MVDW definitions in the test sets. Overall the average error of every training or test set is lower than 1% (0.45−0.89%). For practical purposes, it is also important, however, to minimize maximum errors and error spreads. Here, the error spread is defined as the difference between the maximum positive and negative errors. It can be seen from Table 2 that the unsigned maximum errors by the revised density definition are smaller than those of the density and the MVDW definitions (revised density: 1.72−3.38%; density: 2.23−8.44%; and MVDW: 5.27−9.32%) and that the error spreads of the revised density definition are narrower (revised density: 2.50−5.90%; density: 2.90−10.85%; and MVDW: 4.39−10.47%). Overall, the revised density definition can achieve a more consistent performance with lower average errors, maximum errors, and error spreads than those of the density and MVDW definitions in both the training and test sets.

We next consider the correlation between the reaction field energies computed with the revised density definition and those with the SES definition for test set PDB2. As shown in Figure 2, all data fall on the diagonal line with almost no scatter. Indeed, the correlation coefficient is 0.999979, the slope is 0.999505, the offset is 5.30848, and the root-mean-square (rms) relative deviation is 0.57%. In contrast, the correlation between the density definition and the SES definition is 0.999607, the slope is 1.01631, the offset is 11.1488, and the rms relative deviation is 1.79%. The correlation between the MVDW and the SES definitions is 0.999876, the slope is 1.01112, the offset is −15.306, and the rms relative deviation is 2.17%.

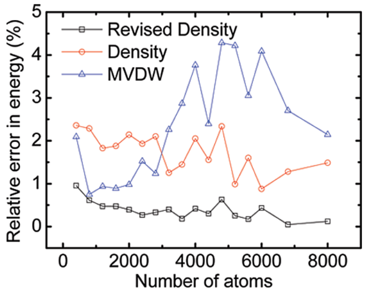


*Figure 3.* Size dependence of relative errors in reaction field energies for the revised density (square), the density (circle), and the MVDW (triangle) definitions. To improve clarity, the numbers of atoms are binned with a width of 400, and the error is averaged for each bin before plotting.

**Transferability Considerations.** While the overall relative error in reaction field energy should be as small as possible, it is also important that the error is both system and structure independent. We investigated two aspects in this regard: errors with different system sizes and different conformations (loosely packed/unfolded or folded conformations). To study the transferability of the new method, we calculated the errors versus number of atoms for the PDB2 set. Figure 3 shows that the error range of the revised density definition does not change much as the number of atoms increases. This confirms that the performance of the revised density and the density definitions are mostly size independent, though the revised density definition clearly agrees better with the SES definition. On the other hand, the size-dependence effect of the MVDW definition is obvious; the error increases as the number of atoms increases.

Besides the size dependence, we also studied the conformation dependence of the revised density definition. We unfolded two proteins (SHG/ENH) and two peptides (helix/HPN2) with high-temperature (500 K) molecular dynamics simulations and analyzed the correlation between energy errors and conformations for about 250−500 collected snapshots for each molecule. Figure 4 shows that the conformation dependence of the revised density definition is lower, i.e., the energy errors are uniform over the different conformations, shown as different rmsd's. It is worth pointing out that in the SHG and ENH test sets, the reaction field energies of the revised density definition are similar to those of the density definition. In the helix and HPN2 test sets, however, the results of the revised density definition are much better than both the density and MVDW definitions. This demonstrates the more consistent performance of the revised density definition over different sized molecules.

**Convergence and Stability of Atomic Electrostatic Forces.** Given the overall improved agreement in the computed reaction field energies between the revised density

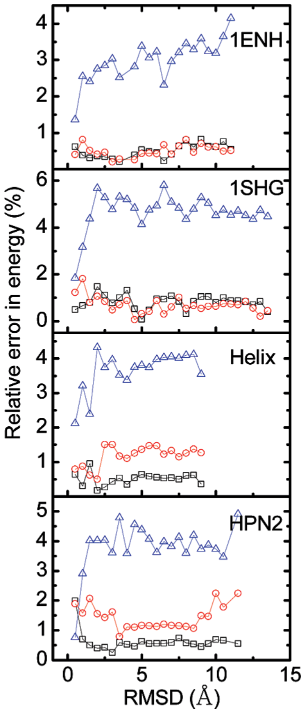


***Figure 4.*** Conformation dependence of relative unsigned errors in reaction field energies for the revised density (square), the density (circle), and the MVDW (triangle) definitions. To improve clarity, the rmsd's are binned with a width of 0.5 Å, and the error is averaged for each bin before plotting.

and the SES definitions, we studied the convergence and the numerical stability of atomic electrostatic forces, which are important for stable dynamics simulations. The convergence quality of atomic electrostatic forces is measured by the correlation and the rmsd of atomic forces at a typical coarse grid spacing (1/2 Å) with respect to those at a fine grid spacing (1/8 Å). The numerical stability of atomic forces is measured by the average standard deviations of individual atomic forces computed with different finite-difference grid origins. Here 64 different finite-difference grid origins were used to analyze the numerical uncertainty of different methods. Two small systems, the native helix (from test set helix) and the native hairpin (from test set HPN2) structures, were selected in this analysis because they can be processed with the finest tested grid spacing, 1/8 Å, on our local computer servers.

*Consistency of the Two Methods.* While the two surface definitions are clearly different, we have tried to optimize the revised density definition so that the reaction field energies computed with the two surface definitions agree as much as possible. Thus, we expect the correlation of atomic forces computed with the two surface definitions to be reasonably well, at least at the finest grid spacing tested, 1/8 Å. The correlations between the two sets of forces are shown for the two tested molecules in Figure 5. The reaction field (RF) forces are shown on the left, with the correlation coefficients of 0.98730 for the helix and 0.98912 for the hairpin, indicating that the RF forces by SES are reasonably well reproduced by the revised density function. Similarly, the right panel of Figure 5 plots the correlations of dielectric boundary (DB) forces between the two methods for the two tested molecules. The correlation coefficients are 0.95137 for the helix and 0.95848 for the hairpin. Due to the physical nature of the dielectric boundary forces, which act on the molecular surface, they are more sensitive to the exact location of molecular surface. Therefore, the correlation of DB forces is, in principle, lower than that of the RF forces between the two surface definitions.

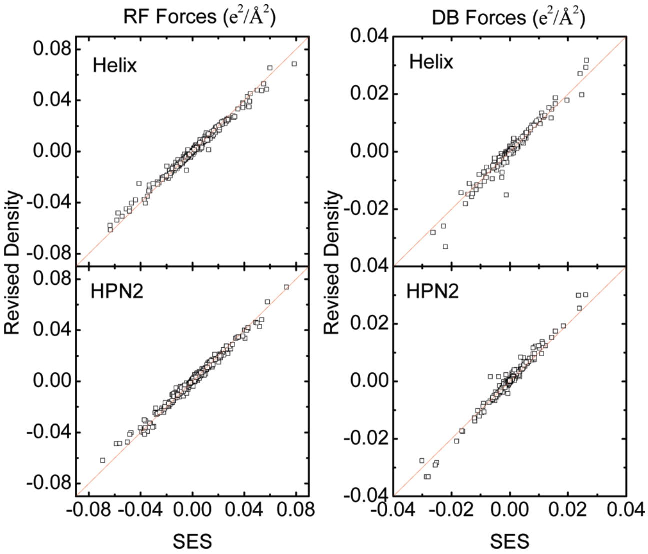


***Figure 5.*** Left: Correlation between reaction field (RF) forces computed with the revised density definition and those computed with the SES definition. Right: Correlation between dielectric boundary (DB) forces computed with the revised density definition and those computed with the SES definition. All forces are computed with a grid spacing of 1/8 Å.

*Convergence of Forces.* We next analyzed the convergence of forces at typical coarse grid spacing (1/2 Å) for biomolecular simulations for both the revised density and the SES definitions. The convergence measures show that the revised density definition converges faster than the SES definition in the numerical calculation of RF forces (Table 3) and DB forces (Table 4). The correlations of RF forces for the revised density function are 0.99411 for the helix and 0.98917 for the hairpin, respectively, while the corresponding correlations for the SES definition are 0.96625 and 0.91612, respectively. The rms deviations of RF forces computed with the revised density function are a factor about two smaller than the corresponding rms deviations with the SES definition. The convergence comparison for the DB forces is similar. The correlations of DB forces computed with the revised density function are 0.93229 for the helix and 0.95588 for the hairpin, respectively, while the corresponding correlations computed with the SES definition are 0.90205 and 0.84782, respectively. The rms deviations of DB forces computed with the revised density function are a factor of about two smaller than the corresponding rms deviations with the SES definition.

*Numerical Stability of Forces.* Finally, we analyzed the numerical stability of numerical electrostatic forces for both the revised density and the SES definitions. Our stability measure shows that the revised density definition is more stable than that of the SES definition in the numerical calculation of RF and DB forces (Tables 3 and 4); reductions

***Table 3.*** Convergence and Numerical Stability of Reaction Field forces ($e^2$/ Å$^2$) for the Tested Helix and Hairpin at the Coarse Grid Spacing of 1/2 Å[a]

| | | revised density | | | SES | | |
|---|---|---|---|---|---|---|---|
| | 1/h | CC | rmsd (×10$^{-4}$) | $\sigma$ (×10$^{-4}$) | CC | rmsd (×10$^{-4}$) | $\sigma$ (×10$^{-4}$) |
| helix | 2 | 0.99411 | 12.7 | 3.40 | 0.96625 | 22.5 | 5.04 |
| | 8 | NA | NA | 0.192 | NA | NA | 0.274 |
| hairpin | 2 | 0.98917 | 18.5 | 5.40 | 0.91612 | 39.5 | 7.59 |
| | 8 | NA | NA | 0.238 | NA | NA | 0.326 |

[a] SES: solvent excluded surface definition. CC: correlation coefficient between the forces at 1/2 and 1/8 Å. rmsd: rms deviation between the forces at 1/2 and 1/8 Å. $\sigma$: average standard deviation of individual atomic forces.

***Table 4.*** Convergence and Numerical Stability of Dielectric Boundary Forces ($e^2$/ Å$^2$) for the Tested Helix and Hairpin at the Coarse Grid Spacing of 1/2 Å[a]

| | | revised density | | | SES | | |
|---|---|---|---|---|---|---|---|
| | 1/h | CC | rmsd (×10$^{-4}$) | $\sigma$ (×10$^{-4}$) | CC | rmsd (×10$^{-4}$) | $\sigma$ (×10$^{-4}$) |
| helix | 2 | 0.93229 | 16.2 | 7.08 | 0.90205 | 22.4 | 15.9 |
| | 8 | NA | NA | 1.58 | NA | NA | 2.27 |
| hairpin | 2 | 0.95588 | 12.8 | 12.5 | 0.84782 | 33.0 | 24.9 |
| | 8 | NA | NA | 2.64 | NA | NA | 2.77 |

[a] See Table 3 for more information.

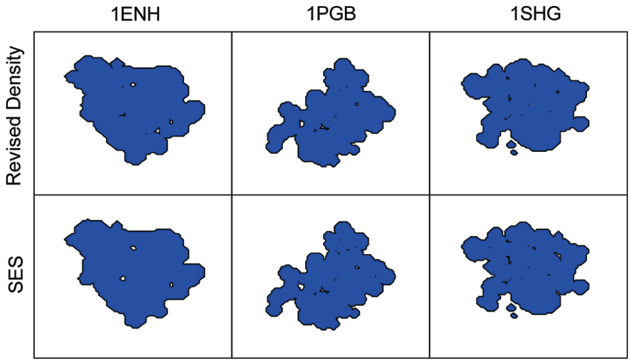


***Figure 6.*** Molecular surface determined by the revised density and SES definitions for folded conformations of 1ENH, 1PGB, and 1SHG, respectively.

in the standard deviations by a factor of about two were observed when the revised density definition was used for both types of forces.

**Reproduction of Molecular Surface.** Note that we have solely relied on reaction field energies in the optimization of the revised density definition. This is reasonable because the measure based on reaction field energies is quite sensitive to the exact locations of molecular surface. Nevertheless, we have also studied the quality of the revised density definition in reproducing the molecular surfaces with the SES definition. Figure 6 shows the molecular surfaces for folded 1ENH, 1PGB, and 1SHG. It can be seen that both definitions are quite consistent for these folded proteins. Figure S-1 in the Supporting Information shows the molecular surfaces for unfolded 1ENH, 1PGB, and 1SHG with the two definitions, respectively. Here slightly more discrepancy can be observed in the computed molecular surface. Overall, the molecular surface of both folded and unfolded conformations in the SES definition can be reasonably reproduced by the revised density function.

**Reproduction of Surface Electrostatic Potential.** We further analyzed the surface electrostatic potential of several selected proteins with the revised density definition. Figure 7 shows the surface potentials for 1PGB, 1ENH, and 1SHG calculated with the revised density function and the SES definition, respectively. Clearly, the two surface potential maps agree very well for the three selected proteins. For the larger p53 tetramer with and without DNA, the agreement is also excellent, as shown in Figure S-2 in the Supporting Information.

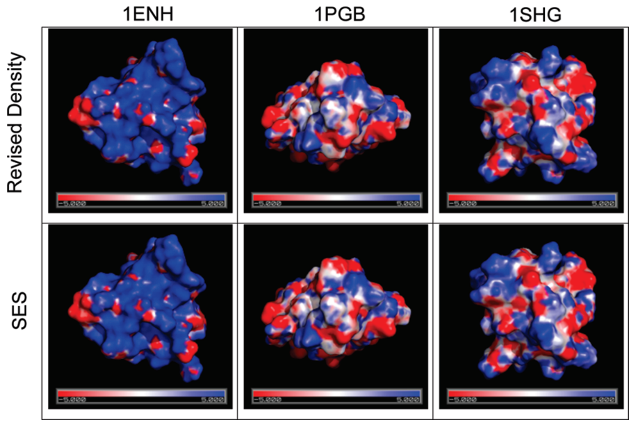


***Figure 7.*** Surface electrostatic potential (in kT/mol-e) for protein 1ENH, 1PGB, and 1SHG, respectively (from left to right). Upper panel: computed with the revised density definition. Lower panel: computed with the SES definition. Potential is visualized in PyMol using a continuous color scale. Blue: positive values; white: zero; and red: negative.

**Limitation of the Density Function Approaches.** A limitation of the revised density definition concerns the use of a "unconventionally" small solvent probe, 0.6 Å, which was optimized to reproduce the electrostatic potential of mean forces of dimers and surface salt bridges on macromolecules in the TIP3P explicit solvent. It is likely that the use of a small probe may result in a more hydrophilic interior in loosely packed macromolecules. Thus, we also tried to use the traditional solvent probe of 1.4 Å in our training of the revised density function. However, our analysis shows that the revised density definition can perform no better than the MVDW definition, even in the training sets (data not shown).

Thus, the limitation of the method to applications with a smaller solvent probe radii is clearly an issue that needs further considerations, to make it a more general numerical procedure for molecular surface calculations. Nevertheless, since our ultimate motivation in the study was to reproduce as close as possible the explicit solvent energetics, and it is unclear how quantitative loosely packed the molecular interior would be without detailed explicit solvent simulations, we delay further improvement to a future study. In addition, it is worth highlighting the intrinsic difference between all density function definitions and the SES definition to represent the molecular surface. Indeed, the multibody re-entry region in the SES definition is very difficult, if not possible, to reproduce with any additive density function approach.

## Concluding Remarks

A revised density function is developed to define the molecular surface for dielectric assignment in the numerical PB methods. The density function is optimized through an exhaustive search using a benchmark data set of reaction field energies, computed with the solvent excluded surface for a wide-range of proteins and peptides in both folded and unfolded conformations. A uniformly low average unsigned error of less than 1% in reaction field energies can be achieved for both the training and test sets. The maximum errors and error spreads are also smaller than the modified van der Waals and the original density function definitions, upon which the new density function was developed. We further studied the transferability of the revised density function. Our data show that the average unsigned errors change little as the molecular size increases, confirming the size independence of the method. We also studied the conformation dependence of the revised density function with unfolded conformations of two proteins and two peptides. It turns out that there is little conformation dependence in the revised density function. Next, we studied the convergence and the stability of numerical electrostatic forces computed with the revised density function. The analysis shows that the revised density function can improve both the convergence and numerical stability by a factor of two over the solvent excluded surface. The analysis of computed molecular surfaces shows that the revised density function can well reproduce the solvent excluded surfaces for the three tested proteins in both folded and unfolded conformations. Finally, we also analyzed the performance of the revised density function in reproducing the surface electrostatic potential. The analysis shows that the agreements are very good for different sized proteins, whether it is neutral or highly charged.

Nevertheless, it is worth pointing out the limitation of the current development documented here. First, it is important to highlight that the numerical instability of the dielectric boundary forces is still higher than that of the reaction field forces. This is, in part, due to the use of grid-independent surface polarization charges in the computation of the reaction field energy and forces, which improves the convergence and stability in the computed energies and forces. In contrast, the dielectric boundary forces were computed directly with the finite-difference grid potential without further treatment. This is, in part, responsible for their larger error and uncertainty. We are actively working to improve the convergence and the stability of dielectric boundary forces by utilizing the grid-independent surface polarization charges, which would further enhance their numerical performance when combined with the revised density function. In addition, partition of dielectric boundary force elements on the solvent/surface interface to nearby atoms is clearly the next natural step before the method can be applied to routine molecular dynamics simulations. We are actively working on both issues in this group.

Finally, our analysis shows that the density approximation does not work well for “conventional” solvent probes of 1.4 Å. Fortunately, our prior comparative analysis indicates that the numerical PB methods with a probe radius of 0.6 Å reproduces the TIP3P solvent best on the tested conformations and systems. It is likely that the use of a small probe may result in a more hydrophilic interior in loosely packed macromolecules. This apparent dilemma between the molecular interior being too hydrophilic and the best agreement with TIP3P solvent points to the limitation of the hard-sphere-based strategies to define the dielectric model for numerical PB methods. To overcome this difficulty, we may revise the molecular surface calculation into a two-step procedure. First, we shall use a large solvent probe to calculate solvent accessibility of each atom. Of course, this procedure should be optimized to be consistent with solvent accessibility simulated in explicit solvent models. For the solvent inaccessible atoms, we will augment their VDW radii with an optimized amount that would effectively fill the molecular interior, resulting in a hydrophobic interior. Second, we will still use the density function optimized with a solvent probe of 0.6 Å to compute the re-entry region among the solvent-accessible atoms. The proposed strategy is still within the hard-sphere strategies but may resolve the dilemma discussed to a certain degree.

**Acknowledgment.** This work is supported in part by NIH (GM069620).

**Supporting Information Available:** The larger p53 tetramer with and without DNA. The molecular surfaces for unfolded 1ENH, 1PGB, and 1SHG with the two definitions, respectively. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Davis, M. E.; McCammon, J. A. *Chem. Rev.* **1990**, *90*, 509.

(2) Gilson, M. K. *Curr. Opin. Struct. Biol.* **1995**, *5*, 216.

(3) Honig, B.; Nicholls, A. *Science* **1995**, *268*, 1144.

(4) Sharp, K. A. *Curr. Opin. Struct. Biol.* **1994**, *4*, 234.

(5) Bashford, D.; Case, D. A. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129.

(6) Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161.

(7) Roux, B.; Simonson, T. *Biophys. Chem.* **1999**, *78*, 1.

(8) Baker, N. A. *Curr. Opin. Struct. Biol.* **2005**, *15*, 137.

(9) Chen, J. H.; Im, W. P.; Brooks, C. L. *J. Am. Chem. Soc.* **2006**, *128*, 3728.

(10) Feig, M.; Chocholousova, J.; Tanizaki, S. *Theor. Chem. Acc.* **2006**, *116*, 194.

(11) Im, W.; Chen, J. H.; Brooks, C. L. In *Peptide Solvation and H-Bonds*; Elsevier Academic Press Inc: San Diego, CA, 2006; Vol. 72, p 173.

(12) Koehl, P. *Curr. Opin. Struct. Biol.* **2006**, *16*, 142.

(13) Lu, B. Z.; Zhou, Y. C.; Holst, M. J.; McCammon, J. A. *Communications in Computational Physics* **2008**, *3*, 973.

(14) Wang, J.; Tan, C. H.; Tan, Y. H.; Lu, Q.; Luo, R. *Communications in Computational Physics* **2008**, *3*, 1010.

(15) Davis, M. E.; McCammon, J. A. *J. Comput. Chem.* **1989**, *10*, 386.

(16) Klapper, I.; Hagstrom, R.; Fine, R.; Sharp, K.; Honig, B. *Proteins: Struct., Funct., Genet.* **1986**, *1*, 47.

(17) Luty, B. A.; Davis, M. E.; McCammon, J. A. *J. Comput. Chem.* **1992**, *13*, 1114.

(18) Nicholls, A.; Honig, B. *J. Comput. Chem.* **1991**, *12*, 435.

(19) Forsten, K. E.; Kozack, R. E.; Lauffenburger, D. A.; Subramaniam, S. *J. Phys. Chem.* **1994**, *98*, 5580.

(20) Holst, M.; Saied, F. *J. Comput. Chem.* **1993**, *14*, 105.

(21) Im, W.; Beglov, D.; Roux, B. *Comput. Phys. Commun.* **1998**, *111*, 59.

(22) Rocchia, W.; Alexov, E.; Honig, B. *J. Phys. Chem. B* **2001**, *105*, 6507.

(23) Bashford, D. *Lecture Notes in Computer Science* **1997**, *1343*, 233.

(24) Gilson, M. K.; Sharp, K. A.; Honig, B. H. *J. Comput. Chem.* **1988**, *9*, 327.

(25) Luo, R.; David, L.; Gilson, M. K. *J. Comput. Chem.* **2002**, *23*, 1244.

(26) Rocchia, W.; Sridharan, S.; Nicholls, A.; Alexov, E.; Chiabrera, A.; Honig, B. *J. Comput. Chem.* **2002**, *23*, 128.

(27) Madura, J. D.; Briggs, J. M.; Wade, R. C.; Davis, M. E.; Luty, B. A.; Ilin, A.; Antosiewicz, J.; Gilson, M. K.; Bagheri, B.; Scott, L. R.; McCammon, J. A. *Comput. Phys. Commun.* **1995**, *91*, 57.

(28) Nicholls, A.; Sharp, K. A.; Honig, B. *Proteins: Struct., Funct., Genet.* **1991**, *11*, 281.

(29) Wang, J.; Cai, Q.; Li, Z. L.; Zhao, H. K.; Luo, R. *Chem. Phys. Lett.* **2009**, *468*, 112.

(30) Cai, Q.; Wang, J.; Zhao, H. K.; Luo, R. *J. Chem. Phys.* **2009**, *130*, 145101.

(31) Cai, Q.; Hsieh, M. J.; Wang, J.; Luo, R. *J. Chem. Theory Comput.* **2010**, *6*, 203.

(32) Wang, J.; Luo, R. *J. Comput. Chem.* **2010**, DOI: 10.1002/jcc.21456.

(33) Baker, N.; Holst, M.; Wang, F. *J. Comput. Chem.* **2000**, *21*, 1343.

(34) Cortis, C. M.; Friesner, R. A. *J. Comput. Chem.* **1997**, *18*, 1591.

(35) Holst, M.; Baker, N.; Wang, F. *J. Comput. Chem.* **2000**, *21*, 1319.

(36) Chen, L.; Holst, M. J.; Xu, J. C. *SIAM J. Numerical Anal.* **2007**, *45*, 2298.

(37) Shestakov, A. I.; Milovich, J. L.; Noy, A. *J. Colloid Interface Sci.* **2002**, *247*, 62.

(38) Xie, D.; Zhou, S. *BIT Numerical Mathematics* **2007**, *47*, 853.

(39) Hoshi, H.; Sakurai, M.; Inoue, Y.; Chujo, R. *J. Chem. Phys.* **1987**, *87*, 1107.

(40) Miertus, S.; Scrocco, E.; Tomasi, J. *Chem. Phys.* **1981**, *55*, 117.

(41) Zauhar, R. J.; Morgan, R. S. *J. Comput. Chem.* **1988**, *9*, 171.

(42) Juffer, A. H.; Botta, E. F. F.; Vankeulen, B. A. M.; Vanderploeg, A.; Berendsen, H. J. C. *J. Comput. Phys.* **1991**, *97*, 144.

(43) Rashin, A. A. *J. Phys. Chem.* **1990**, *94*, 1725.

(44) Yoon, B. J.; Lenhoff, A. M. *J. Comput. Chem.* **1990**, *11*, 1080.

(45) Bharadwaj, R.; Windemuth, A.; Sridharan, S.; Honig, B.; Nicholls, A. *J. Comput. Chem.* **1995**, *16*, 898.

(46) Purisima, E. O.; Nilar, S. H. *J. Comput. Chem.* **1995**, *16*, 681.

(47) Zhou, H. X. *Biophys. J.* **1993**, *65*, 955.

(48) Liang, J.; Subramaniam, S. *Biophys. J.* **1997**, *73*, 1830.

(49) Vorobjev, Y. N.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 569.

(50) Boschitsch, A. H.; Fenley, M. O.; Zhou, H. X. *J. Phys. Chem. B* **2002**, *106*, 2741.

(51) Lu, B. Z.; Cheng, X. L.; Huang, J. F.; McCammon, J. A. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 19314.

(52) Totrov, M.; Abagyan, R. *Biopolymers* **2001**, *60*, 124.

(53) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127.

(54) Onufriev, A.; Bashford, D.; Case, D. A. *J. Phys. Chem. B* **2000**, *104*, 3712.

(55) Tsui, V.; Case, D. A. *J. Am. Chem. Soc.* **2000**, *122*, 2489.

(56) Onufriev, A.; Case, D. A.; Bashford, D. *J. Comput. Chem.* **2002**, *23*, 1297.

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

(58) Feig, M.; Im, W.; Brooks, C. L. *J. Chem. Phys.* **2004**, *120*, 903.

(59) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins: Struct., Funct., Bioinf.* **2004**, *55*, 383.

(60) Tanizaki, S.; Feig, M. *J. Chem. Phys.* **2005**, 122.

(61) Sigalov, G.; Scheffel, P.; Onufriev, A. *J. Chem. Phys.* **2005**, 122.

(62) Mongan, J.; Simmerling, C.; McCammon, J. A.; Case, D. A.; Onufriev, A. *J. Chem. Theory Comput.* **2007**, *3*, 156.

(63) Grant, J. A.; Pickup, B. T.; Sykes, M. J.; Kitchen, C. A.; Nicholls, A. *Phys. Chem. Chem. Phys.* **2007**, *9*, 4913.

(64) Davis, M. E. *J. Chem. Phys.* **1994**, *100*, 5149.

(65) Luo, R.; Moult, J.; Gilson, M. K. *J. Phys. Chem. B* **1997**, *101*, 11226.

(66) Li, X. F.; Hassan, S. A.; Mehler, E. L. *Proteins: Struct., Funct., Bioinf.* **2005**, *60*, 464.

(67) Eisenhaber, F.; Argos, P. *J. Comput. Chem.* **1993**, *14*, 1272.

(68) You, T.; Bashford, D. *J. Comput. Chem.* **1995**, *16*, 743.

(69) Zauhar, R. J.; Morgan, R. S. *J. Comput. Chem.* **1990**, *11*, 603.

(70) Swanson, J. M. J.; Mongan, J.; McCammon, J. A. *J. Phys. Chem. B* **2005**, *109*, 14769.

(71) Tan, C. H.; Yang, L. J.; Luo, R. *J. Phys. Chem. B* **2006**, *110*, 18680.

(72) Wang, J.; Tan, C.; Chanco, E.; Luo, R. *Phys. Chem. Chem. Phys.* **2010**, *12*, In press.

(73) Lu, Q.; Luo, R. *J. Chem. Phys.* **2003**, *119*, 11035.

(74) Chocholousova, J.; Feig, M. *J. Comput. Chem.* **2006**, *27*, 719.

(75) Alexov, E. *Proteins: Struct., Funct., Bioinf.* **2003**, *50*, 94.

(76) Davis, M. E.; McCammon, J. A. *J. Comput. Chem.* **1991**, *12*, 909.

(77) Grant, J. A.; Pickup, B. T.; Nicholls, A. *J. Comput. Chem.* **2001**, *22*, 608.

(78) Grant, J. A.; Pickup, B. T. *J. Phys. Chem.* **1995**, *99*, 3503.

(79) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes in C: The Art of Scientific Computing*; Cambridge University Press: New York, NY, 1992.

(80) Stoer, J.; Bulirsch, R. *Introduction to Numerical Analysis*, 3rd ed.; Springer-Verlag: New York, NY,2002.

(81) Gilson, M. K.; Davis, M. E.; Luty, B. A.; McCammon, J. A. *J. Phys. Chem.* **1993**, *97*, 3591.

(82) Osher, S.; Fedkiw, R. P. *Level set methods and dynamic implicit surfaces*; Springer: New York, NY, 2003.

(83) Sethian, J. A. *Level set methods and fast marching methods: evolving interfaces in computational geometry, fluid mechanics, computer vision, and materials science*; Cambridge University Press: New York, NY, 1999.

(84) Landau, L. D.; Lifshitz, E. M.; Pitaevskii, L. P. *Electrodynamics of Continuous Media*, 2nd ed.; Butterworth-Heinemann: Oxford, 1993.

(85) Ye, X.; Wang, J.; Luo, R., submitted.

(86) Sharp, K. A.; Honig, B. *J. Phys. Chem.* **1990**, *94*, 7684.

(87) Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 265.

(88) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668.

CT900318U

JCTC Journal of Chemical Theory and Computation

# Colored-Noise Thermostats à la Carte

Michele Ceriotti,*,[†] Giovanni Bussi,[‡] and Michele Parrinello[†]

*Computational Science, Department of Chemistry and Applied Biosciences, ETH Zürich, USI Campus, Via Giuseppe Buffi 13, CH-6900 Lugano, Switzerland and Università di Modena e Reggio Emilia and INFM-CNR-S3, via Campi 213/A, 41100 Modena, Italy*

Received October 23, 2009

**Abstract:** Recently, we have shown how a colored-noise Langevin equation can be used in the context of molecular dynamics as a tool to obtain dynamical trajectories whose properties are tailored to display desired sampling features. In the present paper, after having reviewed some analytical results for the stochastic differential equations forming the basis of our approach, we describe in detail the implementation of the generalized Langevin equation thermostat and the fitting procedure used to obtain optimal parameters. We also discuss the simulation of nuclear quantum effects and demonstrate that by carefully choosing parameters one can successfully model strongly anharmonic solids such as neon. For the reader's convenience, a library of thermostat parameters and some demonstrative code can be downloaded from an online repository.

## 1. Introduction

Stochastic differential equations (SDE) have been used to model the time evolution of processes characterized by random behavior in fields as diverse as physics and economics. In particular, the Langevin equation (LE) has been regularly applied in the study of Brownian motion and used extensively in molecular dynamics (MD) computer simulations as a convenient and efficient tool to obtain trajectories which sample the constant-temperature, canonical ensemble.[1,2]

In its original form, the Langevin equation is based on the assumption of instantaneous system−bath interactions, which corresponds to the values of the random force being uncorrelated at different times. A non-Markovian, generalized version of the LE arises in the context of Mori−Zwanzig theory.[3,4] In this theory, if one considers a harmonic system coupled with a harmonic bath, it is possible to integrate out the degrees of freedom of the bath. This leaves one with a linear stochastic equation where both the friction and the noise have a finite memory. The conventional LE is recovered in the limit of a clear separation between the characteristic time scale of the system's dynamics and that of the system−bath interaction.

This class of non-Markovian SDEs has been extensively used to model the dynamics of open systems interacting with a physically relevant bath (see, e.g., refs 5–7). Instead, our recent works[8,9] have used colored(correlated)-noise SDEs as a device to sample efficiently statistical distributions in molecular-dynamics (MD) simulations. These works aimed to show how a stochastic thermostat suitable for Car−Parrinello-like dynamics[8] could be constructed, and how nuclear quantum effects can be included in a large class of problems at a fraction of the cost of path-integrals calculations.[9] In these applications the real dynamics is lost, and one focuses only on the efficient calculation of static ensemble averages.

In this paper we discuss the practical implementation of the generalized Langevin equation (GLE) thermostat that we used in the two cases mentioned above. We also provide the reader with the analytical and numerical tools needed to construct SDEs tailored to their own sampling needs. Throughout we take advantage of the dimensional reduction scheme, which allows one to exploit the equivalence between non-Markovian dynamics and Markovian dynamics in higher dimensionality. In doing this, we supplement the physical coordinates with additional degrees of freedom,[4] whose equations of motion are taken as linear, so as to simplify the formalism and analytical derivations.

In the Appendices we recall some of the properties of multidimensional stochastic processes,[5,10–12] which are useful

* Corresponding author e-mail: michele.ceriotti@phys.chem.ethz.ch.

[†] ETH Zürich.

[‡] Università di Modena e Reggio Emilia and INFM-CNR-S3.

10.1021/ct900563s © 2010 American Chemical Society
Published on Web 03/01/2010

to our discussion, and present a short comparison of the GLE thermostat and the widely used massive Nosé−Hoover chains.[13−16] A simple FORTRAN90 code implementing our method to the dynamics of a harmonic oscillator and a library of optimized thermostat parameters can be downloaded from an online repository.[17]

## 2. Generalized Langevin Thermostat

**2.1. Markovian and Non-Markovian Formulations.** The Langevin equation for a particle with position $q$ and momentum $p$, subject to a potential $V(q)$, can be written as

$$\begin{aligned}\dot{q} &= p\\ \dot{p} &= -V'(q) - a_{pp}p + b_{pp}\xi(t)\end{aligned} \tag{1}$$

where $\xi(t)$ represents an uncorrelated, Gaussian-distributed random force with unitary variance and zero mean [$\langle\xi\rangle = 0$, $\langle\xi(t)\xi(0)\rangle = \delta(t)$]. Here and in what follows we use mass-scaled coordinates. Furthermore, for consistency, the friction coefficient (usually denoted by $\gamma$) is here given the symbol $a_{pp}$, while $b_{pp}$ is the intensity of the random force. In this notation, the fluctuation−dissipation theorem (FDT) reads $b_{pp}^2 = 2a_{pp}k_BT$. If this relation holds, the dynamics generated by eq 1 will sample the canonical ensemble at temperature $T$.[4,18]

As explained in the Introduction, in order to bypass the complexity of dealing with a non-Markovian formulation directly, we supplement the system with $n$ additional degrees of freedom $\mathbf{s} = \{s_i\}$, which are linearly coupled to the physical momentum and between themselves. The resulting SDE can be cast into the compact form

$$\begin{aligned}\dot{q} &= p\\ \begin{pmatrix}\dot{p}\\ \dot{\mathbf{s}}\end{pmatrix} &= \begin{pmatrix}-V'(q)\\ 0\end{pmatrix} - \begin{pmatrix}a_{pp} & \mathbf{a}_p^T\\ \bar{\mathbf{a}}_p & \mathbf{A}\end{pmatrix}\begin{pmatrix}p\\ \mathbf{s}\end{pmatrix} + \begin{pmatrix}b_{pp} & \mathbf{b}_p^T\\ \bar{\mathbf{b}}_p & \mathbf{B}\end{pmatrix}\begin{pmatrix}\xi\end{pmatrix}\end{aligned} \tag{2}$$

Here, $\xi$ is a vector of $n + 1$ uncorrelated Gaussian random numbers with $\langle\xi_i(t)\xi_j(0)\rangle = \delta_{ij}\delta(t)$. Clearly, eq 1 is recovered when $n = 0$. For a harmonic potential $V(q) = \omega^2q^2/2$, eqs 2 are linear and an Ornstein−Uhlenbeck (OU) process is recovered whose time propagation can be evaluated analytically. In the nonlinear case one can use the Trotter decomposition to split the dynamics into a linear part, which evolves the $(p, \mathbf{s})$ momenta, and a nonlinear part, which evolves Hamilton's equations.[19] This is facilitated by the fact that the dynamics of $(p, \mathbf{s})$ alone is linear, and its exact finite-time propagator can be analytically evaluated (see section 2.5).

Here and in the rest of the paper, we adopt the same notation introduced in ref 9 to distinguish between matrices acting on the full state vector $\mathbf{x} = (q, p, \mathbf{s})^T$ or on parts of it as illustrated below:

$$\begin{array}{c|c|c|c|} & q & p & \mathbf{s}\\ \hline q & m_{qq} & m_{qp} & \mathbf{m}_q^T\\ \hline p & \bar{m}_{qp} & m_{pp} & \mathbf{m}_p^T\\ \hline \mathbf{s} & \bar{\mathbf{m}}_q & \bar{\mathbf{m}}_p & \mathbf{M}\\ \hline \end{array}\quad \}\mathbf{M}_p \quad \}\mathbf{M}_{qp} \tag{3}$$

The Markovian dynamical eqs 2 are equivalent to a non-Markovian process for the physical variables only. This is best seen by first considering the evolution of the $(p, \mathbf{s})$ variables in the free-particle analogue of eqs 2. The additional degrees of freedom $\mathbf{s}$ can be integrated away, and one is left with (see ref 4 and Appendix A)

$$\dot{p} = -\int_{-\infty}^{t} K(t - s)p(s)\mathrm{d}s + \zeta(t) \tag{4}$$

where the memory kernel $K(t)$ is related to the elements of $\mathbf{A}_p$ by

$$K(t) = 2a_{pp}\delta(t) - \mathbf{a}_p^T e^{-|t|\mathbf{A}}\bar{\mathbf{a}}_p \tag{5}$$

On the basis of the fact that the free-particle dynamics of $(p, \mathbf{s})$ is an OU process, one also finds than the relationship between the static covariance matrix $\mathbf{C}_p = \langle(p, \mathbf{s})^T(p, \mathbf{s})\rangle$, the drift matrix $\mathbf{A}_p$, and the diffusion matrix $\mathbf{B}_p$ is given by

$$\mathbf{A}_p\mathbf{C}_p + \mathbf{C}_p\mathbf{A}_p^T = \mathbf{B}_p\mathbf{B}_p^T \tag{6}$$

Note the remarkable formal analogy between eq 6 and the equations for the orthogonality constraints in Car−Parrinello dynamics, see, e.g., ref 20. In Appendix A we show that setting $\mathbf{C}_p = k_BT$ is sufficient to satisfy the FDT. In this case, eq 6 fixes $\mathbf{B}_p$ once $\mathbf{A}_p$ is given. FDT also implies that the colored-noise autocorrelation function $H(t) = \langle\zeta(t)\zeta(0)\rangle$ is equal to $k_BTK(t)$, whereas the more complex relation between $K(t)$ and $H(t)$, valid in the general case, is reported in eq 27.

Since there is no explicit coupling between the position $q$ and the additional momenta $\mathbf{s}$, one can check that exactly the same dimensional reduction can be performed in the case of an arbitrary potential coupling $p$ and $q$ and that eqs 2 correspond to the non-Markovian process

$$\begin{aligned}\dot{q} &= p\\ \dot{p} &= -\frac{\partial V}{\partial q} - \int_{-\infty}^{t} K(t - s)p(s)\mathrm{d}s + \zeta(t)\end{aligned} \tag{7}$$

In the memory kernel eq 5, $\mathbf{A}$ can be chosen to be a general real matrix and can have complex eigenvalues, provided they have a positive real part. This results in a $K(t)$ that is a linear combination of exponentially damped oscillations. Therefore, a vast class of non-Markovian dynamics can be represented by Markovian equations such as eqs 2.

**2.2. Exact Solution in the Harmonic Limit.** The thermostats typically used in MD simulations have a few parameters that are chosen by trial and error. A thermostat based on eqs 2 depends on a much larger number of parameters, and hence the fitting procedure is more complex. It is therefore important to find ways to compute a priori analytical estimates so as to guide the tuning of the thermostat.

To this end, we examine the harmonic oscillator, which is commonly used to model physical and chemical systems. By choosing $V(q) = \omega^2q^2/2$ the force term in eqs 2 becomes linear and the dynamics of $\mathbf{x} = (q, p, \mathbf{s})^T$ is the OU process $\dot{\mathbf{x}} = -\mathbf{A}_{qp}\mathbf{x} + \mathbf{B}_{qp}\xi$. In eqs 2 the $\mathbf{s}$ degrees of freedom are coupled to the momentum only. Therefore, most of the additional entries in $\mathbf{A}_{qp}$ and $\mathbf{B}_{qp}$ are zero, and the equations for $\mathbf{x}$ read

$$\begin{pmatrix} \dot{q} \\ \dot{p} \\ \dot{\mathbf{s}} \end{pmatrix} = -\begin{pmatrix} 0 & -1 & \mathbf{0} \\ \omega^2 & a_{pp} & \mathbf{a}_p^T \\ \mathbf{0} & \bar{\mathbf{a}}_p & \mathbf{A} \end{pmatrix} \begin{pmatrix} q \\ p \\ \mathbf{s} \end{pmatrix} + \begin{pmatrix} 0 & 0 & \mathbf{0} \\ 0 & \multicolumn{2}{c}{\mathbf{B}_p} \\ \mathbf{0} & & \end{pmatrix} \begin{pmatrix} 0 \\ \xi \end{pmatrix} \tag{8}$$

The exact finite-time propagator for eq 8 can be computed, and so it is possible to obtain any ensemble average or time-correlation function analytically. Of course, one is most interested in the expectation values of the physical variables $q$ and $p$. In particular, one can obtain the fluctuations $\langle q^2 \rangle$ and $\langle p^2 \rangle$ and correlation functions of the form $\langle q^2(t) q^2(0) \rangle$, which can be used to measure the coupling between the thermostat and the system. The resulting expressions are simple to evaluate but lengthy, and we refer the reader to Appendix B for their explicit form.

One can envisage, using the estimates computed for an oscillator of frequency $\omega$, to predict and hence optimize the response of a normal mode of a similar frequency in the system being studied. Furthermore, thanks to the properties of eq 8, one does not need to perform a normal-modes analysis to turn this idea into a practical method. Consider indeed a perfect harmonic crystal and apply an independent instance of the GLE thermostat to the three Cartesian coordinates of each atom. It is easy to see that since eq 8 is linear and contains Gaussian noise, the thermostatted equations of motion are invariant under any orthogonal transformation of the coordinates. Therefore, the resulting dynamics can be described on the basis of the normal modes just as in ordinary Hamiltonian lattice dynamics. As a consequence, each phonon will respond independently as a 1-D oscillator with its own characteristic frequency. Thus, to tune the GLE thermostat, one only needs the analytical results in the one-dimensional case, evaluated as a function of $\omega$. The parameters can then be optimized for a number of different purposes, based solely on minimal information on the vibrational spectrum of the system under investigation, without any knowledge of the eigenmodes of the phonons.

The invariance properties of the GLE thermostat lead to additional advantages. For instance, we can contrast its behavior with that of Nosé–Hoover (NH) chains, based on equations which are quadratic in $p$ (see Appendix C). As a consequence of the nonlinearity, the efficiency of an NH chains thermostat for a multidimensional oscillator depends on the orientation of the eigenmodes relative to the Cartesian axes, an artifact which is absent in our case.

Having set the background, we now turn to the description of the various applications of eqs 2.

**2.3. Efficient Canonical Sampling.** We first discuss the design of a GLE which can optimally sample phase space. In this case, the target stationary distribution is the canonical ensemble, so the equations of motion need to satisfy the detailed-balance condition. Still, there is a great deal of freedom available in the choice of the autocorrelation kernel or, equivalently, in the choice of the $\mathbf{A}_p$ matrix. These free parameters can be used to optimize the sampling efficiency. To this end, we must first define an appropriate merit function. Standard choices are the autocorrelation times of the potential and total energy ($V$ and $H$, respectively):

$$\begin{aligned} \tau_V &= \frac{1}{\langle V^2 \rangle - \langle V \rangle^2} \int_0^\infty \langle (V(t) - \langle V \rangle)(V(0) - \langle V \rangle) \rangle \mathrm{d}t \\ \tau_H &= \frac{1}{\langle H^2 \rangle - \langle H \rangle^2} \int_0^\infty \langle (H(t) - \langle H \rangle)(H(0) - \langle H \rangle) \rangle \mathrm{d}t \end{aligned} \tag{9}$$

In the harmonic case, these can be readily computed in terms of correlation times of $q^2$ and $p^2$ (see Appendix B) and will depend on $\mathbf{A}_p$ and the oscillator's frequency $\omega$. For example, one easily finds that in the white-noise limit with no additional degrees of freedom as in eq 1

$$\tau_H(\omega) = \frac{1}{a_{pp}} + \frac{a_{pp}}{4\omega^2}, \qquad \tau_V(\omega) = \frac{1}{2a_{pp}} + \frac{a_{pp}}{2\omega^2} \tag{10}$$

Both response times are constant in the high-frequency limit and increase quadratically in the low-frequency extreme of the spectrum. For a given frequency one can choose $a_{pp}$ so as to minimize the correlation time, thus enhancing sampling. It should be noted that eqs 10 contain a "trivial" dependence on $\omega$, as one expects that sampling a normal mode would require at least a time on the order of its vibrational period. One can thus define a renormalized $\kappa(\omega) = [\tau(\omega)\omega]^{-1}$ as a measure of the efficiency of the coupling. In the white-noise case, $\kappa = 1$ for the optimally coupled frequency ($\omega_H = a_{pp}/2$ and $\omega_V = a_{pp}$, respectively) and decreases linearly for lower and higher values of $\omega$.

While this result in itself provides a guide to choose a good value of the friction coefficient in conventional (white-noise) Langevin dynamics, we can enhance the value of $\kappa(\omega)$ over a broader frequency range by using a colored-noise SDE. If we want to obtain canonical sampling, the FDT has to hold, so that $\mathbf{C}_p = k_\mathrm{B}T$. We therefore consider the entries of $\mathbf{A}_p$ as the only independent parameters, since $\mathbf{B}_p$ is then determined by eq 6.

In practice, we set up a fitting procedure in which we choose a set of frequencies $\omega_i$ distributed over a broad range ($\omega_\mathrm{min}$, $\omega_\mathrm{max}$). For an initial guess for the thermostat matrix $\mathbf{A}_p$ we compute $\kappa(\omega)$ for each of these frequencies. We then vary $\mathbf{A}_p$ so as to optimize $\min_i \kappa(\omega_i)$ and aim at a sampling efficiency on the range ($\omega_\mathrm{min}$, $\omega_\mathrm{max}$) which is as high and frequency independent as possible. We will discuss this fitting procedure in more detail in section 3.

In Figure 1 we compare the optimized $\kappa(\omega)$ for different frequency ranges and number of additional degrees of freedom. We find empirically that $\kappa(\omega) = 1$ is the best result which can be attained and that nearly optimal efficiency can be reached over a very broad range of frequencies. This constant efficiency decreases slightly as the fitted range is extended, regardless of the number $n$ of $s_i$ employed. For a given frequency range, however, increasing $n$ has the effect of making the response flatter.

Clearly this scheme will work optimally in harmonic or quasi-harmonic systems, and anharmonicity will introduce deviations from the predicted behavior. In the extreme case of diffusive systems such as liquids, one has to ask the question of how much diffusion will be affected by the thermostat, especially since in an overdamped LE equation the diffusive modes are considerably slowed down (see, e.g., ref 21). To estimate the impact of the thermostat on the

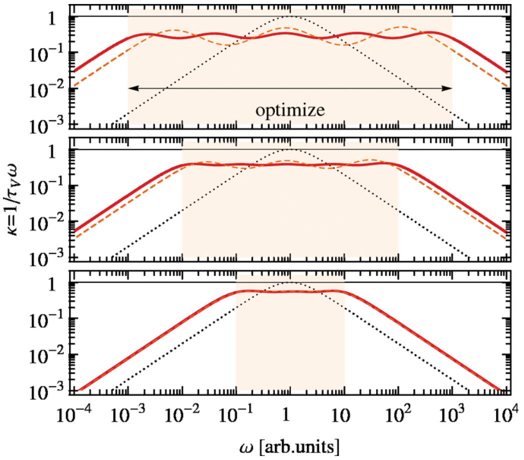


***Figure 1.*** Sampling efficiency as estimated from eq 9 for a harmonic oscillator, plotted as a function of the frequency $\omega$. The $\kappa(\omega)$ curve for a white-noise Langevin thermostat optimized for $\omega = 1$ (black, dotted lines, eq 10) is contrasted with those for a set of optimized GLE thermostats. The panels, from bottom to top, contain the results fitted, respectively, over a frequency range spanning 2, 4, and 6 orders of magnitudes around $\omega = 1$. Dark, continuous lines correspond to matrices with $n = 4$, and dashed, lighter lines correspond to $n = 2$. The GLE curves correspond to the sets of parameters kv_4−2, kv_2−2, kv_4−4, kv_2−2, kv_4−6, kv_2−6, which can be downloaded from an online repository.[17]

diffusion, we define the free-particle diffusion coefficient $D^*$ as that calculated switching off the physical forces. Its value when a GLE thermostat is used is

$$\begin{aligned}\frac{mD^*}{k_BT} &= \frac{1}{\langle p^2\rangle}\int_0^\infty \langle p(t)p(0)\rangle\,\mathrm{d}t \\ &= [\mathbf{A}_p^{-1}]_{pp} = (a_{pp} - \mathbf{a}_p^T\mathbf{A}^{-1}\bar{\mathbf{a}}_p)^{-1}\end{aligned} \tag{11}$$

where we assumed the FDT to hold. In practical cases, if an estimate of the unthermostated (intrinsic) diffusion coefficient $D$ is available, one should choose the matrix $\mathbf{A}_p$ in such a way that $D^* \gg D$, so that the thermostat will not behave as an additional bottleneck for diffusion. Equation 11 has the interesting consequence that $D^*$ can be enhanced either by reducing the overall strength of the noise, as in white-noise LE, or by carefully balancing the terms in the denominator of eq 11.

We found empirically that for an $\mathbf{A}_p$ matrix fitted to harmonic modes over the frequency range ($\omega_{\min}$, $\omega_{\max}$) the diffusion coefficient computed by eq 11 is $D^* \approx k_BT/m\omega_{\min}$. This latter expression gives a useful recipe for choosing the minimal frequency to be considered when fitting a GLE thermostat for a system whose diffusion coefficient can be roughly estimated.

**2.4. Frequency-Dependent Thermostatting.** The ability to control the strength of the thermostat−system coupling as a function of the frequency, demonstrated above, points quite naturally at more sophisticated applications. For instance, one can apply two thermostats with distinct target temperatures and different efficiencies $\kappa(\omega)$ (see Figure 2). Obviously, such a simulation is not an equilibrium one, since energy is systematically injected in some modes and removed from others, but leads to a steady state that has useful properties. Indeed, the normal modes will couple differently to the two thermostat, so that the effective temperature of each mode can be controlled as a function of $\omega$. This two-thermostats example is just an instance of a broader class of stochastic processes, for which the FDT is violated. In general, we can relax the assumption that $\mathbf{C}_p = k_BT$, and for a given drift matrix we can choose a $\mathbf{B}_p$, which is suitable to our purpose.

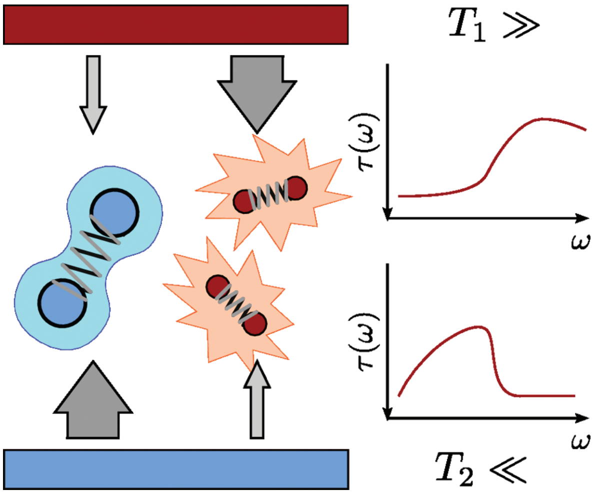


***Figure 2.*** Cartoon representing a two-thermostat setup, which we take as the simplest example of a stochastic process violating the fluctuation−dissipation theorem. If the relaxation time versus frequency curves for the two thermostats are different, a steady state will be reached in which normal modes corresponding to different frequencies will equilibrate at different effective temperatures.

Returning to the harmonic oscillator case, one can solve exactly the dynamics for a given choice of $\mathbf{A}_p$, $\mathbf{B}_p$, and frequency $\omega$. The resulting dynamics is performed in the $(n+2)$-dimensional space defined by the variables $(q, p, \mathbf{s})$ according to eq 8. For a compact notation, we used the full matrices $\mathbf{A}_{qp}$ and $\mathbf{B}_{qp}$. The full $\mathbf{C}_{qp}(\omega)$, which defines the stationary distribution in the steady state, can be computed solving an equation analogous to eq 6

$$\mathbf{A}_{qp}\mathbf{C}_{qp} + \mathbf{C}_{qp}\mathbf{A}_{qp}^T = \mathbf{B}_{qp}\mathbf{B}_{qp}^T \tag{12}$$

One can tune the free parameters ($\mathbf{A}_p$ and $\mathbf{B}_p$) so as to make the $c_{qq}(\omega)$ and $c_{pp}(\omega)$ elements of the extended covariance matrix as close as possible to the desired target functions $\langle q^2\rangle(\omega)$ and $\langle p^2\rangle(\omega)$.

In a previous paper[9] we applied this method to obtain $\langle q^2\rangle(\omega)$ and $\langle p^2\rangle(\omega)$ in agreement with the values appropriate for a quantum harmonic oscillator and obtained a good approximation to the quantum-corrected structural properties in quasi-harmonic systems. Many other applications can be envisaged, which take advantage of frequency-dependent thermostatting. For instance, one could use this technique in accelerated sampling methods,[22–24] which work by artificially heating the low-frequency modes while keeping the other modes at the correct temperature.

**2.5. Implementation.** The implementation of a GLE thermostat in molecular-dynamics simulations is straightforward. Here, we consider the case of a velocity−Verlet integrator, which updates positions and momenta by a time step $\Delta t$, according to the scheme:

$$\begin{aligned} p &\leftarrow p - V'(q)\Delta t/2 \\ q &\leftarrow q + p\Delta t \\ p &\leftarrow p - V'(q)\Delta t/2 \end{aligned} \tag{13}$$

Equations 13 can be obtained using Trotter splitting in a Liouville operator formalism.[25] In the same spirit one can introduce our GLE thermostat by performing two free-particle steps by $\Delta t/2$ on the $(p, \mathbf{s})$ variables:[19]

$$\begin{aligned} (p,\mathbf{s}) &\leftarrow \mathscr{P}[(p,\mathbf{s}),\Delta t/2] \\ p &\leftarrow p - V'(q)\Delta t/2 \\ q &\leftarrow q + p\Delta t \\ p &\leftarrow p - V'(q)\Delta t/2 \\ (p,\mathbf{s}) &\leftarrow \mathscr{P}[(p,\mathbf{s}),\Delta t/2] \end{aligned} \tag{14}$$

At variance with thermostats based on second-order equations of motion such as Nosé−Hoover, where a multiple time-step approach is required to obtain accurate trajectories,[26,27] this free-particle step can be performed without introducing additional sampling errors. The exact finite-time propagator for $(p, \mathbf{s})$ reads

$$\mathscr{P}[(p,\mathbf{s}),\Delta t]^T = \mathbf{T}(\Delta t)(p,\mathbf{s})^T + \mathbf{S}(\Delta t)\boldsymbol{\xi}^T \tag{15}$$

where $\boldsymbol{\xi}$ is a vector of $n + 1$ uncorrelated Gaussian numbers and the matrices $\mathbf{T}$ and $\mathbf{S}$ can be computed once at the beginning of the simulation and for all degrees of freedom.[10,28] The relations between $\mathbf{T}$, $\mathbf{S}$, $\mathbf{A}_p$, $\mathbf{C}_p$, and $\Delta t$ read

$$\mathbf{T} = e^{-\Delta t\mathbf{A}_p}, \qquad \mathbf{S}\mathbf{S}^T = \mathbf{C}_p - e^{-\Delta t\mathbf{A}_p}\mathbf{C}_p e^{-\Delta t\mathbf{A}_p^T}$$

It is worth pointing out that when FDT holds, the canonical distribution is invariant under the action of eq 15, whatever the size of the time step. A useful consequence of this property is that in the rare cases where applying eq 15 introduces a significant overhead over the force calculation, the thermostat can be applied every $m$ steps of dynamics using a stride of $m\Delta t$. This will change the trajectory but does not affect the accuracy of sampling.

The velocity−Verlet algorithm (eq 13) introduces finite-$\Delta t$ errors, whose effect needs to be monitored. In microcanonical simulations, this is routinely done by checking conservation of the total energy $H$. Following the work of Bussi et al.[29] we introduce a conserved quantity $\tilde{H}$, which can be used for the same purpose

$$\tilde{H} = H - \sum_i \Delta K_i \tag{16}$$

where $\Delta K_i$ is the change in kinetic energy due to the action of the thermostat at the $i$th time step and the sum is extended over the past trajectory. In cases where the FDT holds, such as that described in section 2.3, the drift of the effective energy quantitatively measures the violation of detailed balance induced by the velocity−Verlet step, similarly to refs 19 and 29. In the cases where the FDT does not hold, such as the frequency-dependent thermostatting described in section 2.4, the conservation of this quantity just measures the accuracy of the integration, similarly to refs 30 and 31.

## 3. Fitting of Colored-Noise Parameters

A key feature of our approach resides in the possibility to optimize the performance of the thermostat based on analytical estimates, making the method effectively parameterless. Such optimization, however, is not trivial, even if computationally inexpensive. The relationship between $\mathbf{A}_p$, $\mathbf{B}_p$, and the correlation properties of the resulting trajectory is highly nonlinear. Furthermore, we found empirically that many local minima exist which greatly hinder the optimization process. With these difficulties in mind, we provide a downloadable library of fitted parameters[17] which can be adapted to most of the foreseeable applications, according to the prescriptions given in section 3.4. Details about the fitting procedure are given in the following three sections.

**3.1. Parameterization of GLE Matrices.** A number of constraints must be enforced on the drift and diffusion matrices in order to guarantee that the resulting SDE is well behaved. It is therefore important to find a representation of the matrices such that during fitting these conditions are automatically enforced and that the parameters space is efficiently explored. A first condition, required to yield a memory kernel with exponential decay, is that all the eigenvalues of $\mathbf{A}_p$ must have positive real part. A second requirement is that the kernel $K(\omega)$ is positive for all real $\omega$. This ensures that the stochastic process will be consistent with the second law of thermodynamics.[32]

Finding the general conditions for $\mathbf{A}_p$ to satisfy this second constraint is not simple. However, we can state that a sufficient condition for $K(\omega) > 0$ is that $\mathbf{A}_p + \mathbf{A}_p^T$ is positive definite. For simplicity we shall assume such a positivity condition to hold, since we found empirically that this modest loss of generality does not significantly affect the accuracy or the flexibility of the fit. Moreover, in the case of canonical sampling, $\mathbf{A}_p + \mathbf{A}_p^T > 0$ is also required in order to obtain a real diffusion matrix, since $\mathbf{B}_p\mathbf{B}_p^T = k_BT(\mathbf{A}_p + \mathbf{A}_p^T)$ according to eq 6.

One would like to find a convenient parametrization, which automatically enforces these constraints. This is best done by writing $\mathbf{A}_p = \mathbf{A}_p^{(S)} + \mathbf{A}_p^{(A)}$, the sum of a symmetric and antisymmetric part. Since any orthogonal transform of the $\mathbf{s}$ degrees of freedom would not change the dynamics (see Appendix A), one can assume without loss of generality that the $\mathbf{A}^{(S)}$ block in $\mathbf{A}_p^{(S)}$ is diagonal (see eq 3 for the naming convention). Since in the general case the antisymmetric $\mathbf{A}_p^{(A)}$ does not commute with $\mathbf{A}_p^{(S)}$, we will assume it to be full, while $\mathbf{A}_p^{(S)}$ can be written in the form

$$\mathbf{A}_p^{(S)} = \begin{pmatrix} a & a_1 & a_2 & \dots & a_n \\ a_1 & \alpha_1 & 0 & \dots & 0 \\ a_2 & 0 & \alpha_2 & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ a_n & 0 & 0 & \dots & \alpha_n \end{pmatrix} \tag{17}$$

In order to enforce the positive definiteness, one uses an analytical Cholesky decomposition $\mathbf{A}_p^{(S)} = \mathbf{Q}_p\mathbf{Q}_p^T$ with

$$\mathbf{Q}_p = \begin{pmatrix} q & q_1 & q_2 & \dots & q_n \\ 0 & d_1 & 0 & \dots & 0 \\ 0 & 0 & d_2 & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & d_n \end{pmatrix} \tag{18}$$

and $\alpha_i = d_i^2$, $a_i = d_i q_i$, and $a = q^2 + \Sigma_i q_i^2$. Such a parametrization guarantees that $\mathbf{A}_p$ will generate a dynamics with a stationary probability distribution and requires $2n + 1$ parameters for the symmetric part (the elements of $\mathbf{Q}_p$, eq 18) and $n(n + 1)/2$ for the antisymmetric part $\mathbf{A}_p^{(A)}$. If we want the equilibrium distribution to be the canonical, we must enforce the FDT and $\mathbf{B}_p\mathbf{B}_p^T$ is uniquely determined.

If we aim at a generalized formulation, which allows for frequency-dependent thermalization, there are no constraints on the choice of $\mathbf{B}_p$ other than the fact that both $\mathbf{B}_p\mathbf{B}_p^T$ and the covariance $\mathbf{C}_p$ must be positive definite. Clearly, a real, lower triangular $\mathbf{B}_p$ is the most general parametrization of a positive-definite $\mathbf{B}_p\mathbf{B}_p^T$ and amounts to introducing $(n + 1)(n + 2)/2$ extra parameters. Together with the assumption that $\mathbf{A}_p^{(S)} > 0$, the condition $\mathbf{B}_p\mathbf{B}_p^T > 0$ is sufficient to ensure that the unique symmetric $\mathbf{C}_p$ which satisfies eq 6 is also positive definite.

**3.2. Fitting for Canonical Sampling.** Armed with such a robust and fairly general parametrization, one only needs to define a merit function to be optimized. Again, we first consider the simpler case of canonical sampling. Here, we want to obtain a flat response over a wide, physically relevant frequency range ($\omega_{\min}$, $\omega_{\max}$). We have chosen the form

$$\chi_1 = \left[\sum_i |\log \kappa(\omega_i)|^m\right]^{1/m} \tag{19}$$

where $\omega_i$s are equally spaced on a logarithmic scale over the fitted range. If a large value of $m$ is chosen, the $\omega_i$ which yields the lowest efficiency is weighted more and a flat response curve is obtained. We found empirically that values of $m$ larger than 10 lead to a proliferation of local minima and hinder efficient optimization. To resolve this, one can use the optimal parameters for $m = 2$ as input for further refinement at larger $m$ until convergence is achieved.

This procedure can be modified so as to provide an efficient thermostat which can be used in Car–Parrinello-like dynamics. In this case, the GLE has to act as a low-pass filter in which only the low ionic frequencies are affected and fast electronic modes are not perturbed. To obtain this effect, we compute eq 19 only for the $\omega_i$'s which are smaller than a cutoff frequency $\omega_{CP}$ and introduce an additional term

$$\chi_2 = \left[\sum_{\omega_i > \omega_{CP}} \max\left[\log \kappa(\omega_i) - k \log\frac{\omega_{CP}}{\omega_i}, 0\right]^m\right]^{1/m} \tag{20}$$

$\chi_2$ enforces a steep decrease of $\kappa(\omega)$ above $\omega_{CP}$ with a slope $k$ on a logarithmic scale. Values of $k$ as large as 9 can be used, which guarantee an abrupt drop in thermalization efficiency for the fast modes (see Figure 3).

**3.3. Nonthermal Noise and Quantum Thermostat.** We now discuss the case in which the thermostat is permitted to violate FDT in order to achieve frequency-dependent equilibration. For these applications, one must also fit the fluctuations $c_{pp}(\omega)$ and $c_{qq}(\omega)$ to some target function $\tilde{c}_{pp}$ and $\tilde{c}_{qq}$. We shall not treat the general case but rather investigate the example of the quantum thermostat (ref 9). The procedure followed provides a clear guide for future extensions to different applications.

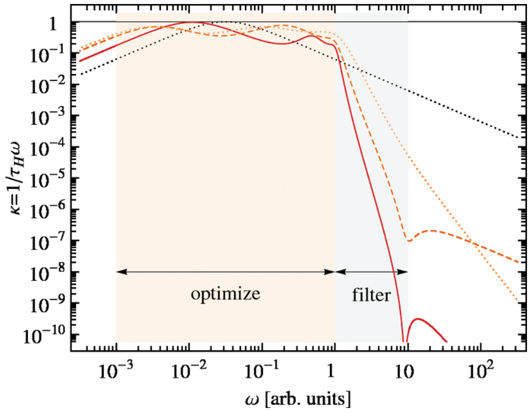


***Figure 3.*** Thermostatting efficiency, as estimated from eq 9, for a colored-noise thermostat optimized for Car–Parrinello dynamics. Sampling efficiency is optimized for $\omega \in (10^{-3}, 1)$, and an abrupt drop in efficiency is enforced for $\omega \in (1, 10)$ using the penalty function eq 20in the fitting. The continuous (dark red) curve corresponds to $k = 9$, the dashed (orange) curve to $k = 6$, and the dotted (light orange) curve to $k = 3$. The $\kappa(\omega)$ curve for a white-noise thermostat centered on the optimized range is also reported for reference (dotted black curve). The three curves correspond to the parameters set, cp-9_4–3, cp-6_4–3 and cp-3_4–3.[17]

In order to reproduce quantum ions effects, one must selectively heat high-frequency phonons for which zero-point energy effects are important without affecting the low-frequency modes which behave classically. The required frequency dependence of the variance for this case is that of a quantum oscillator, i.e., $\tilde{c}_{pp}(\omega) = \omega^2\tilde{c}_{qq}(\omega) = \hbar\omega/2 \coth \hbar\omega/2k_BT$. The $\omega \rightarrow 0$ classical limit can be proved to correspond to two conditions on the elements of the free-particle covariance matrix $\mathbf{C}_p$, namely, $c_{pp} = k_BT$ and $\mathbf{a}_p^T\mathbf{A}^{-1}\mathbf{c}_p = 0$. One could enforce such constraints exactly by considering the entries of $\mathbf{C}_p$ as independent fitting parameters and obtaining the diffusion matrix from eq 6. We found however that this choice makes it difficult to obtain a positive-definite $\mathbf{B}_p\mathbf{B}_p^T$ and that the fitting becomes more complex and inefficient.

As an alternative, we decided to enforce the low-frequency limit with an appropriate penalty function

$$\chi_3 = (c_{pp}/k_BT - 1)^2 + (\mathbf{a}_p^T\mathbf{A}^{-1}\mathbf{c}_p/k_BT)^2 \tag{21}$$

to be optimized together with the sampling efficiency (eq 19) and a term which measures how well the finite-frequency fluctuations were fitted

$$\chi_4 = \left[\sum_i \left|\log\frac{c_{qq}(\omega_i)}{\tilde{c}_{qq}(\omega_i)}\right|^m + \left|\log\frac{c_{pp}(\omega_i)}{\tilde{c}_{pp}(\omega_i)}\right|^m\right]^{1/m} \tag{22}$$

Since the low-frequency limit is already enforced by eq 21, we compute eq 22 on a set of points equally spaced between the maximum frequency $\omega_{\max}$ and one-half of the onset frequency for quantum effects $\omega_q = k_BT/\hbar$.

**3.4. Transferability of Fitted Parameters.** The scheme described in the previous sections allowed us to obtain matrices suitable for all the applications discussed in previous works. Furthermore, it provides a starting point for obtaining matrices which one might deem useful for novel applications. However, the reader is advised that the fitting is still far from being a black-box procedure. It is thus necessary to experiment with a combination of different initial parameters and minimization schemes. We found the downhill simplex method[33] to be particularly effective but resorted to simulated annealing when the optimization got stuck in a local minimum. There is a great deal of arbitrariness in the choice of the terms in eqs 19−22 and in their weighted combination $\chi = \Sigma w_i\chi_i$. To make the procedure even more delicate, we observe that in high-$n$ cases the parameters tend to collapse into "degenerate" minima, where the full dimensionality of the search space is not exploited. This phenomenon can be successfully circumvented by enforcing an even spacing of the eigenvalues of $\mathbf{A}$ over the frequency range of interest and slowly releasing this restraint during the later stages of optimization.

However, the problems mentioned above have no major practical consequences, as the computation of analytical estimates is inexpensive and one can afford a great deal of trial and error during the optimization. Moreover, fitted parameters can be reused, since the optimized parameters can be easily transferred to similar problems because of the scaling properties of the dynamics (eq 8).

In fact, one can see that if the drift and covariance matrices $(\mathbf{A}_p, \mathbf{C}_p)$ lead to the efficiency curves $\kappa(\omega)$ and fluctuations $c_{pp}(\omega)$, the scaled matrices $(\alpha\mathbf{A}_p, \beta\mathbf{C}_p)$ will yield $\kappa(\alpha^{-1}\omega)$ and the fluctuations $\beta c_{pp}(\alpha^{-1}\omega)$. This means that if $\mathbf{A}_p$ is optimized for sampling over the range $(\omega_{\min}, \omega_{\max})$, $\alpha\mathbf{A}_p$ will be optimal over $(\alpha\omega_{\min}, \alpha\omega_{\max})$. We also remark that if $(\mathbf{A}_p, \mathbf{C}_p)$ are fitted to the quantum harmonic oscillator fluctuations at temperature $T$, $(\alpha\mathbf{A}_p, \alpha\mathbf{C}_p)$ will be suitable for temperature $\alpha T$. Care must be taken in this case to ensure that the scaled frequency range still encompasses the whole vibrational spectrum of the system being studied.

## 4. Understanding the Quantum Thermostat

As discussed in ref 9, one must pay a great deal of attention when using a "quantum thermostat" because energy is transferred between modes of different frequency as a consequence of the anharmonic coupling. This is reminiscent of zero-point energy (ZPE) leakage which plagues semiclassical approaches to the computation of nuclear quantum effects.[34,35] In the cases we explored so far, empirical evidence suggests that quasi-harmonic solids can be treated with good accuracy down to temperatures as low as 10% of the Debye temperature $\Theta_D$. Clearly, the ultimate test to assess the accuracy of the method is a comparison with path-integral calculations to be performed on a similar but computationally cheaper model, such as a smaller size box or a simpler force field.

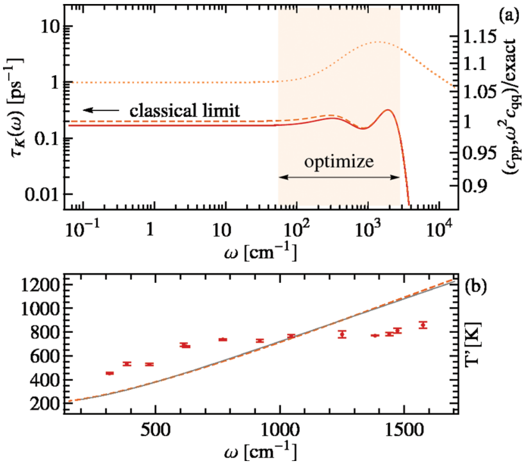


***Figure 4.*** (a) $\omega$ dependence of the kinetic energy correlation time $\tau_k(\omega)$ (light, dotted line) and the ratio of the fitted fluctuations $c_{pp}(\omega)$ (dashed line) and $\omega^2 c_{qq}(\omega)$ (full line) with the exact, quantum-mechanical target function. (b) Normal-mode-projected kinetic temperature for a few selected phonons in diamond. The dashed line is the value expected from the fitted $c_{pp}(\omega)$, while the full line is the exact, quantum-mechanical expectation value for a harmonic oscillator. Calculations have been performed with the parameters qt-20_6_BAD.[17]

One would like however to obtain some qualitative measure of the quality of the fit and gauge the transferability of a given set of parameters. To this end, we first state a couple of empirical rules and then validate them on two fairly different real systems. A first observation is that it is useless to push the fitting of the fluctuations $c_{pp}(\omega)$ and $c_{qq}(\omega)$ to very high accuracy if this comes at the expense of the coupling efficiency. In fact, we would be trading a small, controlled fitting error with a possibly larger, uncontrollable, and system-dependent error stemming from anharmonicity. Second, we observed that in order to contrast more effectively the flow of energy between different phonons, one should try to reduce the correlation time of the kinetic energy $\tau_K$, rather than focus solely on the terms in eq 9, which are better suited to measure sampling efficiency. In fact, a low $\tau_K(\omega)$ corresponds to a slightly overdamped regime, where sampling efficiency is suboptimal but ZPE is enforced more tightly.

To demonstrate these concepts in a real system, we performed some calculations with a Tersoff model of diamond at a temperature $T = 200$ K. At this low temperature, slightly below $0.1\Theta_D$, quantum effects are very strong and we therefore expect to have problems maintaining the large difference in temperature between the stiff and soft phonons. Using a very harmonic system such as diamond is particularly useful, since one can monitor directly the efficiency of the thermostat by projecting the atomic velocities on a selection of normal modes. Hence, a projected kinetic temperature $T'(\omega)$ can be computed and its value checked against the predictions in the harmonic limit in the same spirit as in ref 9. In Figure 4 we report the results with a matrix fitted taking into account only the terms in eqs 21 and 22. Even in a harmonic system such as diamond there are major errors due to ZPE leakage from the high-frequency

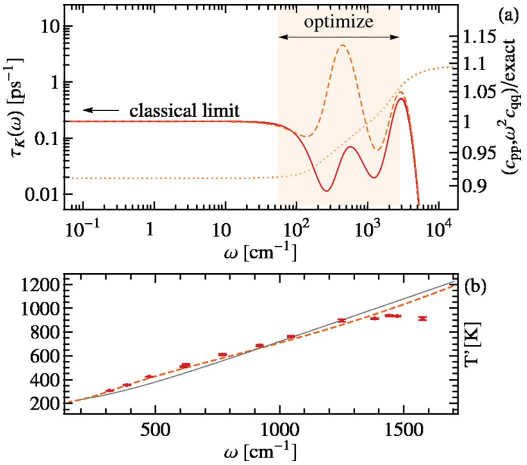


***Figure 5.*** (a) $\omega$ dependence of the kinetic energy correlation time $\tau_K(\omega)$ (light, dotted line) and the ratio of the fitted fluctuations $c_{pp}(\omega)$ (dashed line) and $\omega^2 c_{qq}(\omega)$ (full line) with the exact, quantum-mechanical target function. (b) Normal-mode-projected kinetic temperature for a few selected phonons in diamond. The dashed line is the value expected from the fitted $c_{pp}(\omega)$, while the full line is the exact, quantum-mechanical expectation value for a harmonic oscillator. Calculations have been performed with the parameters qt-20_6.[17]

to the low-frequency modes, which the thermostat compensates only partially. These poor results should be compared with those of Figure 5. Here, we also introduced in the fit a term analogous to 19 to reduce the value of $\tau_K(\omega)$. The projected kinetic temperature now agrees almost perfectly with the analytical predictions $c_{pp}(\omega)$ for most of the modes. The only ones displaying significant deviations are the faster ones, for which the value of $\tau_K(\omega)$ is slightly larger. The $c_{pp}(\omega)$ curve deviates by nearly 10% from the exact, quantum-mechanical expectation value. However, thanks to the more efficient coupling, the errors due to anharmonicities are better compensated, and in actuality, the overall error is much smaller than for the parameters presented in Figure 4.

To test whether these prescriptions work for less harmonic problems, we now turn to a completely different system; namely, the structural properties of solid neon at 20 K. At variance with diamond, quantum-ions effects are less pronounced, but the system is close to its melting temperature and is significantly anharmonic. As shown in Figure 6, the agreement between our results and those of accurate path-integral calculations[36] is almost perfect if the parameters of Figure 5 are used. As expected, large errors are present if qt-20_6_BAD is used. Further improvements on the fitting strategy and the application to strongly anharmonic systems is currently being investigated and will be the subject of further work.

## 5. Conclusions

In this paper we discussed in detail the use of colored-noise dynamics based on Ornstein−Uhlenbeck processes as a tool for performing molecular dynamics. Applications range from enhanced sampling, which we demonstrate in the harmonic limit and will be applied to real systems in forthcoming publications, to thermostats for adiabatically separated problems and frequency-dependent thermalization.

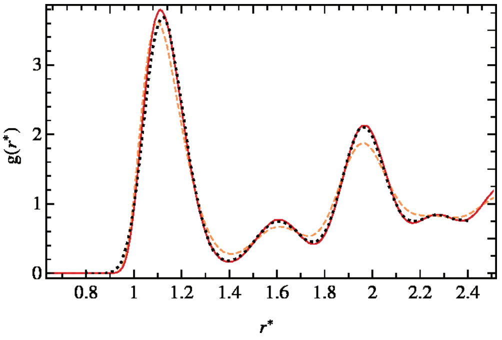


***Figure 6.*** Radial distribution function as computed from fully converged path-integral calculations[36] (black, dotted line) and a quantum-thermostat MD trajectory for a Lennard−Jones model of solid neon at $T = 20$ K. Distances are in reduced units. Full line corresponds to the parameters set qt-20_6 (cf. Figure 5) and lighter, dashed line to the set qt-20_6_BAD (cf. Figure 4).

Our idea exploits the linear nature of the OU stochastic differential equations, which allows one to use the one-dimensional harmonic oscillator as a simple but physically motivated reference model. On the basis of the analytical prediction obtained in that case, we describe a recipe for fitting the thermostat parameters so as to obtain the desired response properties in real systems. The procedure is not simple, and we are considering different approaches to make it more robust and effective. Fortunately, however, fitted matrices can be easily transferred from one system to another. With this in mind we provided an extensive library of optimized parameters,[17] which makes fitting unnecessary for most applications.

We also comment on practical issues concerning the implementation of the generalized-Langevin thermostat in a molecular-dynamics program and its use in applications. In particular, we discuss in detail how one can use colored noise to model nuclear quantum effects.[9] We provide some empirical rules to guide the fitting in this difficult case, and we demonstrate that a normal-mode analysis in a quasi-harmonic system is a valuable tool for assessing the quality of a set of parameters. We believe that further investigation will find many other applications for colored noise in molecular dynamics and in computer simulations of molecular systems in general. As an example, we are currently investigating the use of a zero-temperature, optimal-sampling GLE thermostat in order to perform structural optimization. On similar lines and taking inspiration from “quantum annealing”,[37,38] one can envisage using frequency-dependent thermalization to improve the performance of simulated annealing.

## Appendix A: Memory Kernels for the Non-Markovian Formulation

The connection between the Markovian (eq 2) and non-Markovian (eq 7) formulations of the colored-noise Langevin equation can be understood using techniques similar to those adopted in Mori−Zwanzig theory.[4,11] Let us first consider a very general, multidimensional OU process, where we single out some degrees of freedom (**y**) that we wish to integrate away, leaving only the variables marked as **x**.

$$\left(\frac{\dot{\mathbf{x}}}{\dot{\mathbf{y}}}\right) = -\left(\begin{array}{c|c}\mathbf{A}_{xx} & \mathbf{A}_{xy} \\ \hline \mathbf{A}_{yx} & \mathbf{A}_{yy}\end{array}\right)\left(\frac{\mathrm{x}}{\mathrm{y}}\right) + \left(\frac{\mathbf{B}_{x\xi}}{\mathbf{B}_{y\xi}}\right)\left(\xi\right) \tag{23}$$

Assuming that the dynamics has finite memory, one can safely take $\mathbf{y}(-\infty) = 0$ and the ansatz

$$\mathbf{y}(t) = \int_{-\infty}^{t} e^{-(t-t')\mathbf{A}_{yy}}[-\mathbf{A}_{yx}\mathbf{x}(t') + \mathbf{B}_{y\xi}\xi(t')]\mathrm{d}t' \tag{24}$$

Substituting into eq 23, one sees that **y** can be eliminated from the dynamics of **x** and arrives at

$$\begin{aligned}\dot{\mathbf{x}}(t) &= -\int_{-\infty}^{t}\mathbf{K}(t-t')\mathbf{x}(t')\mathrm{d}t' + \zeta(t)\\ \mathbf{K}(t) &= 2\mathbf{A}_{xx}\delta(t) - \mathbf{A}_{xy}e^{-t\mathbf{A}_{yy}}\mathbf{A}_{yx} \quad (t \geq 0)\\ \zeta(t) &= \mathbf{B}_{x\xi}\xi(t) - \int_{-\infty}^{t}\mathbf{A}_{xy}e^{-(t-t')\mathbf{A}_{yy}}\mathbf{B}_{y\xi}\xi(t')\mathrm{d}t'\end{aligned} \tag{25}$$

One can see that eqs 25 are invariant under any orthogonal transformation of the **y** dynamical variables, meaning that such a transformation leaves the dynamics of the **x**'s unchanged.

The colored noise is better described in terms of its time-correlation function, $\mathbf{H}(t) = \langle\zeta(t)\zeta(0)^T\rangle$. Let us first introduce the symmetric matrix $\mathbf{D} = \mathbf{B}\mathbf{B}^T$, whose parts we shall label using the same scheme used for **A** in eq 23. We shall also need $\mathbf{Z}_{yy} = \int_0^\infty e^{-\mathbf{A}_{yy}t}\mathbf{D}_{yy}e^{-\mathbf{A}_{yy}^T t}\,\mathrm{d}t$. With these definitions in mind, one finds

$$\mathbf{H}(t) = \delta(t)\mathbf{D}_{xx} + \mathbf{A}_{xy}e^{-t\mathbf{A}_{yy}}[\mathbf{Z}_{yy}\mathbf{A}_{xy}^T - \mathbf{D}_{yx}] \quad (t \geq 0) \tag{26}$$

Note that the value of $\mathbf{H}(t)$ for $t < 0$ is determined by the constraint $\mathbf{H}(-t) = \mathbf{H}(t)^T$; the value of $\mathbf{K}(t)$ instead is irrelevant for negative times: we will assume $\mathbf{K}(-t) = \mathbf{K}(t)^T$ to hold, since this will simplify some algebra below.

Let us now switch to the case of the free-particle counterpart of eqs 2, which is relevant to the memory functions entering eqs 7. Here, we want to integrate away all the **s** degrees of freedom, retaining only the momentum $p$. Hence, we can transform eqs 25 and 26 to the less cumbersome form

$$\begin{aligned}K(t) &= 2a_{pp}\delta(t) - \mathbf{a}_p^T e^{-|t|\mathbf{A}}\bar{\mathbf{a}}_p\\ H(t) &= d_{pp}\delta(t) + \mathbf{a}_p^T e^{-|t|\mathbf{A}}[\mathbf{Z}\mathbf{a}_p - \mathbf{d}_p]\end{aligned} \tag{27}$$

This compact notation hides certain relevant property of the memory kernels, which are more apparent when the kernels are written in their Fourier representation. If $\mathbf{D}_p = \mathbf{B}_p\mathbf{B}_p^T$ is transformed according to eq 6. $K(\omega)$ and $H(\omega)$ read

$$\begin{aligned}K(\omega) &= 2a_{pp} - 2\mathbf{a}_p^T\frac{\mathbf{A}}{\mathbf{A}^2+\omega^2}\bar{\mathbf{a}}_p\\ H(\omega) &= K(\omega)\left(c_{pp} - \mathbf{a}_p^T\frac{\mathbf{A}}{\mathbf{A}^2+\omega^2}\mathbf{c}_p\right) +\\ &2\omega^2\left(\mathbf{a}_p^T\frac{1}{\mathbf{A}^2+\omega^2}\mathbf{c}_p\right)\left(1 + \mathbf{a}_p^T\frac{1}{\mathbf{A}^2+\omega^2}\bar{\mathbf{a}}_p\right)\end{aligned} \tag{28}$$

It is seen that the memory functions (hence the dynamical trajectory) are independent of the value of **C**, the covariance of the fictitious degrees of freedom. Moreover, a sufficient condition for the FDT to hold is readily found. By setting $c_{pp} = k_BT$ and $\mathbf{c}_p = 0$, one obtains $H(\omega) = k_BTK(\omega)$, which is precisely the FDT for a non-Markovian Langevin equation. Since the value of **C** is irrelevant we can take $\mathbf{C}_p = k_BT$, which simplifies the algebra and leads to numerically stable trajectories.

## Appendix B: Covariance Matrix and Correlation Times for the Harmonic Oscillator

Given **A** and **C** matrices (the drift term and static covariance for a generic OU process), one can find the diffusion matrix **B** by an expression analogous to eq 6. The same relation can be used to obtain the elements of **C** given the drift and diffusion matrices by solving the linear system. However, the covariance matrix can be computed more efficiently by finding the eigendecomposition of $\mathbf{A} = \mathbf{O}\ \mathrm{diag}(\alpha_i)\ \mathbf{O}^{-1}$ and computing

$$C_{ij} = \sum_{kl}\frac{O_{ik}[\mathbf{O}^{-1}\mathbf{B}\mathbf{B}^T\mathbf{O}^{-1T}]_{kl}O_{jl}}{\alpha_k + \alpha_l} \tag{29}$$

Now, let **x** be the vector describing the trajectory of the OU process. In order to compute $\tau_H$ or $\tau_V$ (eq 9) one needs time-correlation functions of the form $\langle x_i(t)x_j(t)x_k(0)x_l(0)\rangle$. The corresponding, non-normalized integrals

$$\tau_{ijkl} = \int_0^\infty [\langle x_i(t)x_j(t)x_k(0)x_l(0)\rangle - \langle x_ix_j\rangle\langle x_kx_l\rangle]\mathrm{d}t \tag{30}$$

can be computed in terms of the tensorial quantity

$$X_{ijkl} = \sum_{mn}\frac{O_{im}[\mathbf{O}^{-1}\mathbf{C}]_{ml}O_{jn}[\mathbf{O}^{-1}\mathbf{C}]_{nk}}{\alpha_m + \alpha_n} \tag{31}$$

as $\tau_{ijkl} = (1/4)\ (X_{ijkl} + X_{ijlk} + X_{klij} + X_{lkij})$. For example, if we consider the full OU process in the harmonic case, one computes

$$\tau_H = \frac{\omega^4\tau_{qqqq} + 2\omega^2\tau_{qqpp} + \tau_{pppp}}{\omega^4c_{qq}^2 + 2\omega^2c_{qp}^2 + c_{pp}^2}, \qquad \tau_V = \frac{\tau_{qqqq}}{c_{qq}^2} \tag{32}$$

where we use an obvious notation for the indices in $\tau_{ijkl}$.

## Appendix C: Comparison with Nosé−Hoover Chains

The most widespread techniques for canonical sampling in MD are probably white-noise Langevin and Nosé−Hoover chains (NHC). The white-noise Langevin can be considered as a limiting case of the thermostatting method we describe in this work, but NHC is based on a radically different philosophy. It is therefore worth performing a brief comparison between the latter and the GLE thermostat.

In the "massive" version of the NH thermostat,[13,14] each component of the physical momentum is coupled to an additional degree of freedom with a fictitious mass $Q$ by means of a second-order equation of motion. The resulting dynamics ensures that the physically relevant degrees of

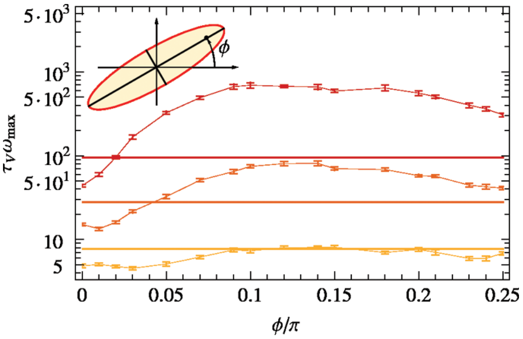


***Figure 7.*** Correlation time for the potential energy of a 2-D harmonic oscillator as a function of the angle between the eigenmodes and the Cartesian axes. $\tau_V$ is computed for different values of the condition number $\omega_{\max}/\omega_{\min}$, from bottom to top 10, 31.6, and 100. Thin lines serve as an aid for the eye, connecting the results obtained in the three cases using a massive NH chains thermostat with four additional degrees of freedom and $Q = k_B T/\omega_{\max}^2$. Error bars are also shown for individual data points. Thick lines correspond to the (constant) result predicted for a GLE thermostat using respectively, the thermostat parameters kv_2−1, centered on $0.32\omega_{\max}$, kv_4−2, centered on $0.18\omega_{\max}$, and kv_4−2 centered on $0.1\omega_{\max}$. The values obtained in actual GLE simulations agree with the predictions within the statistical error bar and are not reported.

freedom will sample the correct, constant-temperature ensemble with the advantage of having deterministic equations of motion and a well-defined conserved quantity. However, in the harmonic case, trajectories are poorly ergodic. This problem can be addressed by coupling the fictitious momentum to a second bath variable with a similar equation of motion. By repeating this process further a "Nosé−Hoover chain" can be formed, which ensures that the dynamics is sufficiently chaotic to achieve efficient sampling.[15,41] The drawback of this approach is that the thermostat equations are quadratic in momenta. It is therefore difficult to obtain analytical predictions for the properties of the dynamics, and the integration of the additional degrees of freedom must be performed with a multiple time-step approach, which makes the thermostat more expensive.

To examine the performances of NHC and GLE, one could envisage comparing the sampling efficiency as defined by the correlation times (eq 9). Obtaining such estimates is not straightforward, not only because the harmonic case cannot be treated analytically but also because in the multidimensional case the properties of the trajectory will not be invariant under an orthogonal transformation of coordinates, as discussed in section 2. The simplest model we can conceive for comparing NHC and GLE is therefore a two-dimensional harmonic oscillator with different vibrational frequencies on the two normal modes and adjustable relative orientations of the eigenvectors with respect to the thermostatted coordinates.

The resulting $\tau_V$ is reported in Figure 7: in the highly anisotropic cases, the efficiency of the NH chains depends dramatically on the orientation of the axes, while for well-conditioned problems is almost constant. The linear stochastic thermostat, on the other hand, has a predictable response, which is completely independent of orthogonal transforms of the coordinates. In the one-dimensional case, or when eigenvectors are perfectly aligned with the axes, NH chains are very efficient for all modes with frequency $\omega < (k_B T/Q)^{1/2}$. One should however consider that in the absence of an exact propagator choosing a small $Q$ implies that integration of the trajectory for the chains will become more expensive.

Obviously, such a simple toy model does not give quantitative information on the behavior in real-life cases, where modes of different frequencies coexist with anharmonicity and diffusive behavior. However, it demonstrates that the colored-noise Langevin thermostat performs almost as well as the axis-aligned NH chains. Furthermore, unlike the NHC, there are no unpredictable failures for anisotropic potentials.

**Acknowledgment.** We thank David Manolopoulos for important suggestions and fruitful discussion, Marcella Iannuzzi for having implemented the GLE thermostat in CP2K,[39] Alessandro Curioni for implementation in CPMD,[40] and Grigorios Pavliotis and Michela Ottobre for discussion and references on stochastic processes. We also acknowledge Stefano Angioletti-Uberti, Paolo Elvati, Hagai Eshet, Kuntal Hazra, Rustam Khaliullin, and Tom Markland for discussion and having preliminarily tested the thermostat in real applications, providing valuable feedback. We are especially in debt to Gareth Tribello, who contibuted to testing and greatly helped us improve the manuscript.

**Note Added after ASAP Publication.** This article was published ASAP on March 1, 2010. Equation 20 has been modified. The correct version was published on March 4, 2010.

## References

(1) Schneider, T.; Stoll, E. *Phys. Rev. B* **1978**, *17*, 1302–1322.

(2) Adelman, S. A.; Brooks, C. L. *J. Phys. Chem.* **1982**, *86*, 1511.

(3) Zwanzig, R. *Phys. Rev.* **1961**, *124*, 983–992.

(4) Zwanzig, R. *Nonequilibrium statistical mechanics*; Oxford University Press: New York, 2001.

(5) Martens, C. C. *J. Chem. Phys.* **2002**, *116*, 2516–2528.

(6) Wang, J.-S. *Phys. Rev. Lett.* **2007**, *99*, 160601.

(7) Kantorovich, L. *Phys. Rev. B* **2008**, *78*, 094304.

(8) Ceriotti, M.; Bussi, G.; Parrinello, M. *Phys. Rev. Lett.* **2009**, *102*, 020601.

(9) Ceriotti, M.; Bussi, G.; Parrinello, M. *Phys. Rev. Lett.* **2009**, *103*, 030603.

(10) Gardiner, C. W. *Handbook of Stochastic Methods*, 3rd ed.; Springer: Berlin, 2003.

(11) Łuczka, J. *Chaos* **2005**, *15*, 026107.

(12) Marchesoni, F.; Grigolini, P. *J. Chem. Phys.* **1983**, *78*, 6287.

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

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

(15) Martyna, G. J.; Tuckerman, M. E.; Klein, M. L. *J. Chem. Phys.* **1992**, *97*, 2635.

(16) Tobias, D. J.; Martyna, G. J.; Klein, M. L. *J. Phys. Chem.* **1993**, *97*, 12959–12966.

(17) GLE4MD; http://gle4md.berlios.de (accessed Jan 12, 2010).

(18) Kubo, R. *Rep. Prog. Phys.* **1966**, *29*, 255–284.

(19) Bussi, G.; Parrinello, M. *Phys. Rev. E* **2007**, *75*, 056707.

(20) Marx, D.; Hutter, J. Ab initio molecular dynamics: Theory and Implementation. In *Modern Methods and Algorithms of Quantum Chemistry Proceedings*, 1st ed.; Grotendorst, J., Ed.; NIC Series: Julich, Germany, 2000; Vol. 1, pp 301−449.

(21) Bussi, G.; Parrinello, M. *Comput. Phys. Commun.* **2008**, *179*, 26.

(22) Rosso, L.; Mináry, P.; Zhu, Z.; Tuckerman, M. E. *J. Chem. Phys.* **2002**, *116*, 4389.

(23) VandeVondele, J.; Rothlisberger, U. *J. Phys. Chem. B* **2002**, *106*, 203–208.

(24) Maragliano, L.; Vanden-Eijnden, E. *Chem. Phys. Lett.* **2006**, *426*, 168–175.

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

(26) Tuckerman, M. E.; Marx, D.; Klein, M. L.; Parrinello, M. *J. Chem. Phys.* **1996**, *104*, 5579–5588.

(27) Jang, S.; Voth, G. A. *J. Chem. Phys.* **1997**, *107*, 9514–9526.

(28) Fox, R. F.; Gatland, I. R.; Roy, R.; Vemuri, G. *Phys. Rev. A* **1988**, *38*, 5938–5940.

(29) Bussi, G.; Donadio, D.; Parrinello, M. *J. Chem. Phys.* **2007**, *126*, 014101.

(30) Bruneval, F.; Donadio, D.; Parrinello, M. *J. Phys. Chem. B* **2007**, *111*, 12219.

(31) Ensing, B.; Nielsen, S. O.; Moore, P. B.; Klein, M. L.; Parrinello, M. *J. Chem. Theory Comput.* **2007**, *3*, 1100–1105.

(32) Ford, G. W.; Lewis, J. T.; O' Connell, R. F. *Phys. Rev. A* **1988**, *37*, 4419–4428.

(33) Nelder, J. A.; Mead, R. *Comp. J.* **1965**, *7*, 308–313.

(34) Alimi, R.; García-Vela, A.; Gerber, R. B. *J. Chem. Phys.* **1992**, *96*, 2034.

(35) Habershon, S.; Manolopoulos, D. E. *J. Chem. Phys.* **2009**, *131*, 244518.

(36) Singer, K.; Smith, W. *Mol. Phys.* **1988**, *64*, 1215–1231.

(37) Lee, Y.-H.; Berne, B. J. *J. Phys. Chem. A* **2000**, *104*, 86.

(38) Santoro, G. E.; Tosatti, E. *J. Phys. A* **2006**, *39*, R393.

(39) The CP2K developers group. CP2K; http://cp2k.berlios.de (accessed Jan 12, 2010).

(40) CPMD; Copyright IBM Corp 1990−2006, Copyright MPI für Festkörperforschung Stuttgart 1997−2001.

(41) Tuckerman, M. E.; Berne, B. J.; Martyna, G. J.; Klein, M. L. *J. Chem. Phys.* **1993**, *99*, 2796–2808.

CT900563S

JCTC Journal of Chemical Theory and Computation

# Accurate Calculation of Hydration Free Energies using Pair-Specific Lennard-Jones Parameters in the CHARMM Drude Polarizable Force Field

Christopher M. Baker,[†] Pedro E. M. Lopes,[†] Xiao Zhu,[†] Benoît Roux,[‡] and Alexander D. MacKerell, Jr.[*,†]

*Department of Pharmaceutical Sciences, School of Pharmacy, University of Maryland, Baltimore, 20 Penn Street, Baltimore, Maryland 21201, and the Department of Biochemistry and Molecular Biology, The University of Chicago, 929 East 57th Street, Chicago, Illinois 60637*

Received October 30, 2009

**Abstract:** Lennard-Jones (LJ) parameters for a variety of model compounds have previously been optimized within the CHARMM Drude polarizable force field to reproduce accurately pure liquid phase thermodynamic properties as well as additional target data. While the polarizable force field resulting from this optimization procedure has been shown to satisfactorily reproduce a wide range of experimental reference data across numerous series of small molecules, a slight but systematic overestimate of the hydration free energies has also been noted. Here, the reproduction of experimental hydration free energies is greatly improved by the introduction of pair-specific LJ parameters between solute heavy atoms and water oxygen atoms that override the standard LJ parameters obtained from combining rules. The changes are small and a systematic protocol is developed for the optimization of pair-specific LJ parameters and applied to the development of pair-specific LJ parameters for alkanes, alcohols and ethers. The resulting parameters not only yield hydration free energies in good agreement with experimental values, but also provide a framework upon which other pair-specific LJ parameters can be added as new compounds are parametrized within the CHARMM Drude polarizable force field. Detailed analysis of the contributions to the hydration free energies reveals that the dispersion interaction is the main source of the systematic errors in the hydration free energies. This information suggests that the systematic error may result from problems with the LJ combining rules and is combined with analysis of the pair-specific LJ parameters obtained in this work to identify a preliminary improved combining rule.

## 1. Introduction

Computer simulations of atomic models are powerful tools that have improved the understanding of many biochemical phenomena, shedding new light on a range of systems from small molecule conformational preferences[1,2] to the dynamics of a complete virus,[3] protein−ligand binding,[4] protein folding,[5] and nucleic acid dynamics.[6] Underpinning such computer simulations is the concept of a force field: a parametrized set of simple differentiable mathematical functions that imitate the quantum mechanical Born−Oppenheimer energy surface and thus allow the calculation of the forces acting on atoms and molecules. Most of the force fields commonly used for the study of biomolecules are based around similar basic concepts,[7] with a series of simplifying approximations introduced to render the simulation of large molecules computationally tractable. One such approximation is that the electrostatic properties of each atom are repre-

[*] Corresponding author phone: (410) 706-7442; fax: (410) 706-5017; e-mail: alex@outerbanks.umaryland.edu.

[†] University of Maryland, Baltimore.

[‡] The University of Chicago.

10.1021/ct9005773 © 2010 American Chemical Society
Published on Web 03/01/2010

sented by a single effective point charge at the site of the nucleus, with energies of electrostatic interactions determined using a Coulomb potential. While this approximation has been both necessary and successful, it neglects the distortion of the electron density around an atom or molecule under the influence of an external field; such models based on fixed effective partial charges ignore the polarizability of the molecule. With increasing computational power available to researchers, the need to use simplified nonpolarizable potential functions in biomolecular simulations is lessened, and simulations based on force fields including an explicit representation of induced polarizability have become feasible.[8–10] Moreover, it is known that there are certain situations in which the omission of polarizability may result in a force field unable to yield accurate results.[7] For example, treatment of the cation−$\pi$ interaction,[11] which is potentially stronger than a conventional hydrogen bond[12] and significant in many biological situations,[13–16] has been shown to require polarizability.[17]

A number of different methods for the explicit inclusion of polarizability into molecular mechanics (MM) force fields are currently being considered.[18] These include methods based on induced point-dipoles,[19,20] classical Drude oscillators,[21] and the fluctuating charge model.[22,23] The CHARMM Drude polarizable force field is an approach based on the classical Drude oscillator model[24] in which polarizability is incorporated via the addition of a "Drude particle" associated with each heavy atom.[21] This auxiliary Drude particle carries a point charge and is attached to its atomic nucleus by a harmonic spring; it is able to relax its position in response to an external field and the relative positions of the fixed charge at the nucleus, and the displacement of the Drude particle then gives rise to an induced dipole moment, accounting explicitly for the polarizability. To date, CHARMM Drude polarizable force field parameters have been developed for a variety of molecules, with a focus on small molecule analogues of the functional groups present within biological macromolecules. Specifically, force field parameters have been obtained for water,[21,25] alkanes,[26] alcohols,[27] aromatics,[28] ethers,[29,30] N-containing aromatic heterocycles,[31] amides,[32] and sulfur-containing compounds.[33] This parametrization has been achieved through extensive fitting to quantum mechanical and experimental reference data using methodologies that have become well-established.[34,35] The resulting parameters have been shown to give satisfactory reproduction of many experimental properties, including liquid and crystal phase thermodynamic properties, liquid phase dielectric constants, dipole moments, interactions with rare gas molecules, and vibrational spectra. However, the force field resulting from this well-established optimization protocol tends to slightly but systematically overestimate the hydration free energies relative to experimental values (i.e., the calculated free energies are too favorable by about 1 kcal/mol).

Clearly, the ability to match experimental hydration free energies accurately (i.e., to within a fraction of a kcal/mol) is highly desirable for a force field that is targeted at the modeling of biomolecular systems. For example, as Xu et al. note, "hydration free energies of amino acids are important because they are directly related to protein folding, protein−protein and protein−membrane interactions."[36] Shirts and Pande further argue that one "cannot expect that calculations performed on more complicated systems, such as those used to compute ligand−protein binding free energies, will be any more accurate than the hydration free energies (or at least the relative hydration free energies) of the respective small constituents."[37] With many of the parameters developed for use in the CHARMM Drude polarizable force field targeted at small molecule analogues of amino acid side chains and drug-like functional groups, these statements alone indicate the importance that should be attached to the accurate reproduction of hydration free energies for all model compounds within the CHARMM Drude polarizable force field.

Accurate calculation of hydration free energies has long been a problem within MM force fields,[37–39] and a variety of approaches have been used in attempts to overcome this problem. Mobley et al. examined the role of atomic partial charges by performing calculations using charge sets derived from increasingly advanced levels of *ab initio* calculation, ultimately concluding that modifying the atomic charges made little difference to the agreement between calculated and experimental hydration free energies.[40] Xu et al. attempted to correct hydration free energies for aromatic groups using an approach in which $\pi$ electron density was represented using a series of non-atom-centered point charges,[41–43] finding that a good reproduction of experimental values could be obtained but, ultimately, that the extra complexity of the model was not justified when comparable improvements could be obtained using a simple reparametrization of the atomic point charges.[36] Having previously identified that additive force fields uniformly "underestimate the solubility of all the (amino acid) side chain analogs",[44] Shirts and Pande[37] came to a similar conclusion. They suggested that the inability of biomolecular force fields to reproduce hydration free energies arose because they were not generally included in the parametrization process. They also concluded that, through careful modification of parameters, it was possible to obtain accurate reproduction of hydration free energies without sacrificing the reproduction of other properties of interest. However, attempts to develop a complete set of parameters for the GROMOS force field based on the simultaneous reproduction of liquid phase thermodynamic properties, free energies of solvation in cyclohexane, and hydration free energies were unsuccessful.[39] The authors concluded that "for almost all functional groups (they) could not find a combination of a charge distribution and a set of van der Waals parameters that would reproduce the free enthalpy of hydration while simultaneously reproducing the density and heat of vaporization of the pure liquid."[39] Instead, they ultimately produced two sets of parameters: one for use in neat liquid simulations and one for use in aqueous phase calculations. Unsurprisingly, the parameter set optimized to reproduce hydration free energies (termed 53A6) was subsequently shown[45] to provide a better reproduction of the hydration free energies of a series of amino acid side chain analogs than did either the AMBER99[46] or OPLS-AA[47,48] models. Both of those models yielded hydration free energies

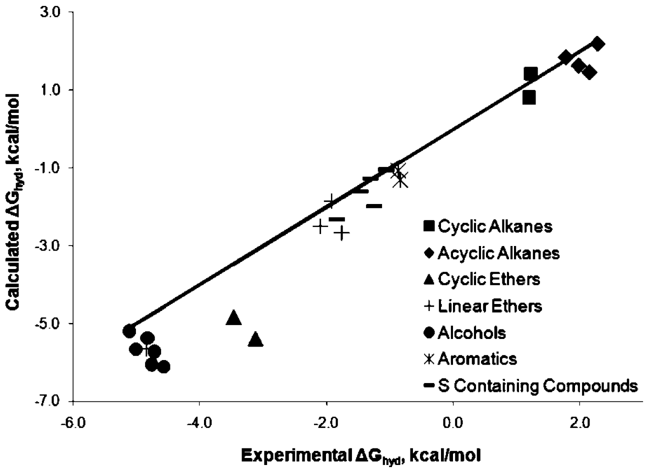


***Figure 1.*** Comparison of experimental hydration free energies with published values calculated using the CHARMM Drude polarizable force field.

that were systematically less favorable than the experimental results. The ability of the 53A6 parameter set to reproduce solvation free energies in a variety of nonaqueous solvents has also been tested, with the parameters yielding results that are generally "in satisfactory agreement with experiment."[49]

One of the most persistently problematic areas for MM force fields has been the accurate representation of the "anomalous" hydration free energies of amines and amides, where the addition of hydrophobic methyl groups results in a more favorable hydration free energy.[50,51] Early additive force fields failed to capture this effect,[52] and attempts to remedy the problem via the inclusion of polarizability also proved unsuccessful.[53,54] Ultimately, the work of Rizzo and Jorgensen[55] and subsequently Chen et al.[56] showed that the errors obtained were due to "nonoptimal parametrization" and that a good reproduction of experimental data could be obtained using a well parametrized additive model with "no need for models with more complex functional forms including explicit polarizability."[55]

Within the CHARMM Drude polarizable force field, hydration free energies calculated using parameters obtained from optimizations primarily targeting the accurate reproduction of pure liquid properties are typically too favorable. Figure 1 shows the relationship between experimental hydration free energies and hydration free energies calculated using the CHARMM Drude polarizable force field taken from the literature, as well as a previously unpublished set of hydration free energies calculated for a series of S containing compounds.[33] While the deviations are small, most are smaller than 1.5 kcal/mol, they are clearly indicative of a systematic problem. There are three points, representing ethane, cyclohexane, and ethane thiol, that lie above the line of perfect correlation, indicating calculated values that are less favorable than the corresponding experimental values. The remaining 22 calculated values, which lie below the line, are more favorable than the corresponding experimental values. For the acyclic alkanes,[26] errors range from 0.07 kcal/mol (4.0%) for ethane to −0.69 kcal/mol (−32.1%) for butane (Table 1). It is also notable that, for the linear alkanes, experimental hydration free energies appear to increase with increasing chain length, while calculated hydration free energies decrease with increasing chain length; the hydration free energies are also too favorable with the alkane parameters[57] for a CHARMM fluctuating charge[58] polarizable force field, and they do not show the decrease in solvation as a function of chain length. For the alcohols,[27] the errors in the calculated values range from −0.09 kcal/mol (2%) for methanol to −1.54 kcal/mol (34%) for butan-2-ol, with the force field again failing to predict correctly the sign of the change in hydration free energy that occurs with increasing chain length (Table 1). Similar results were also obtained for the ethers[30] (Table 1), where all hydration free energies are predicted by the Drude model to be too favorable, with errors ranging from −0.05 kcal/mol (2.6%) for dimethyl ether to −2.22 kcal/mol (71.2%) for tetrahydropyran.

During optimization of Drude parameters for several series of molecules,[27,31] attempts have been made to overcome this problem and provide an accurate reproduction of experimental hydration free energies. These attempts have focused on the use of specific atom−atom Lennard-Jones (LJ) parameters (ie. pair-specific LJ parameters), parameters that can be introduced using the NBFIX option in the CHARMM parameter file thereby overriding the standard LJ parameter combining rules. The use of pair-specific LJ parameters within the Drude model has focused on modifying the interaction between solute atoms and the O atom of the SWM4-NDP[25] polarizable water model and has generally been successful where applied. For example, in the alcohols, the inclusion of pair-specific parameters to modify the interaction between the hydroxyl O and the water O reduced the average error in calculated hydration free energies from 17% to −1%.[27]

Within the CHARMM Drude polarizable force field, the repulsion and dispersion components of the nonbond interaction energy, $E_{\mathrm{LJ}}(r)$, are calculated using a standard LJ potential:

$$E_{\mathrm{LJ}}(r) = \varepsilon\left[\left(\frac{R_{\min}}{r}\right)^{12} - 2\left(\frac{R_{\min}}{r}\right)^{6}\right] \tag{1}$$

where $r$ is the separation between two interacting atoms and $R_{\min}$ and $\varepsilon$ are two empirical parameters, corresponding to the value of $r$ at which $E_{\mathrm{LJ}}(r)$ is a minimum, and the depth of the energy well, respectively. The values of $R_{\min}$ and $\varepsilon$ used to calculate the interaction between two atoms $i$ and $j$ are obtained from individual parameters assigned to each of the two interacting atoms via the following combining rules:

$$R_{\min} = \frac{R_{\min}}{2},i + \frac{R_{\min}}{2},j \tag{2}$$

$$\varepsilon = \sqrt{\varepsilon_i \times \varepsilon_j} \tag{3}$$

When pair-specific LJ parameters are used, however, these standard combining rules are overridden. Values of $R_{\min}$ and $\varepsilon$ for a given atom pair are not calculated from individual contributions arising from each atom but instead are specified directly. This approach allows for the inclusion of pair-specific LJ parameters for any atom pairs of choice, while

***Table 1.*** Hydration Free Energies of Alkanes, Alcohols, and Ethers, All Values in kcal/mol

| molecule | experimental $\Delta G_{\mathrm{hyd}}$ | previously reported Drude $\Delta G_{\mathrm{hyd}}$ | error | without pair-specific LJ parameters $\Delta G_{\mathrm{hyd}}$ | error | with pair-specific LJ parameters $\Delta G_{\mathrm{hyd}}$ | error |
|---|---|---|---|---|---|---|---|
| | | | | Alkanes | | | |
| CPEN | $1.20^a$ | 0.81 ± 0.39 | −0.39 | 0.10 ± 0.05 | −1.10 | 1.16 ± 0.08 | −0.04 |
| CHEX | $1.23^a$ | 1.42 ± 0.21 | 0.19 | 0.44 ± 0.05 | −0.79 | 1.22 ± 0.10 | −0.01 |
| ETHA | $1.77^b$ | 1.84 | 0.07 | 1.64 ± 0.08 | −0.13 | 1.73 ± 0.10 | −0.04 |
| PROP | $1.98^b$ | 1.63 | −0.35 | 1.32 ± 0.04 | −0.66 | 2.04 ± 0.08 | 0.06 |
| BUTA | $2.15^b$ | 1.46 | −0.69 | 1.12 ± 0.12 | −1.04 | 2.08 ± 0.07 | −0.07 |
| IBUT | $2.28^b$ | 2.19 | 0.09 | 1.47 ± 0.08 | −0.81 | 2.25 ± 0.02 | −0.03 |
| NEOP | $2.50^c$ | N/A | N/A | 0.69 ± 0.10 | −1.81 | 2.25 ± 0.12 | −0.26 |
| | | | | average | −0.91 | | −0.06 |
| | | | | Alcohols | | | |
| MEOH | $-5.11^a$ | −5.20 ± 0.19 | −0.09 | −5.20 ± 0.08 | −0.09 | −5.20 ± 0.08 | −0.09 |
| ETOH | $-5.01^a$ | −5.66 ± 0.31 | −0.65 | −5.14 ± 0.07 | −0.13 | −4.85 ± 0.07 | 0.16 |
| PRO2 | $-4.76^a$ | −6.06 ± 0.23 | −1.30 | −5.50 ± 0.05 | −0.74 | −5.41 ± 0.05 | −0.65 |
| BUO2 | $-4.57^c$ | −6.11 ± 0.18 | −1.54 | −5.57 ± 0.08 | −1.00 | −4.21 ± 0.08 | 0.36 |
| PRO1 | $-4.83^a$ | −5.38 ± 0.16 | −0.55 | −5.21 ± 0.08 | −0.38 | −4.96 ± 0.08 | −0.13 |
| BUO1 | $-4.72^a$ | −5.72 ± 0.16 | −1.00 | −5.61 ± 0.09 | −0.89 | −4.74 ± 0.09 | −0.02 |
| | | | | average | −0.54 | | −0.06 |
| | | | | Ethers | | | |
| THF | $-3.47^c$ | −4.80 ± 0.08 | −1.33 | −4.83 ± 0.05 | −1.36 | −3.58 ± 0.05 | −0.11 |
| THP | $-3.12^c$ | −5.34 ± 0.27 | −2.22 | −5.40 ± 0.07 | −2.28 | −3.08 ± 0.10 | 0.04 |
| DEE | $-1.76^c$ | −2.77 ± 0.10 | −1.01 | −2.66 ± 0.15 | −1.76 | −1.83 ± 0.14 | −0.07 |
| DMOE | $-4.84^c$ | −5.61 ± 0.54 | −0.77 | −5.47 ± 0.11 | −0.63 | −5.05 ± 0.13 | −0.21 |
| DME | $-1.92^c$ | −1.97 ± 0.13 | −0.05 | −1.85 ± 0.07 | 0.07 | −1.85 ± 0.06 | 0.07 |
| MEE | $-2.10^d$ | −2.27 ± 0.25 | −0.17 | −2.51 ± 0.08 | −0.41 | −1.78 ± 0.08 | 0.32 |
| | | | | average | −1.06 | | 0.01 |
| | | | | overall average | −0.84 | | −0.03 |

[a] Experimental data from ref 81. [b] Experimental data from ref 50. [c] Experimental data from ref 68. [d] Experimental data from ref 82.

nonbond interactions involving all other atom pairs are calculated using $R_{\min}$ and $\varepsilon$ values obtained via the standard combining rules.

As mentioned above, the pair-specific LJ parameter approach to correcting calculated hydration free energies has been shown to work.[27,31] An objective of the present work is, therefore, to extend this approach to allow for the development of new pair-specific LJ parameters in a more systematic fashion. As an example, consider the case of the alcohols, where alcohol hydration free energies were modified by introducing pair-specific LJ parameters.[27] The alcohol parameters were built upon the alkane parameters with the nonbond parameter optimization focusing on the hydroxyls and adjacent aliphatic moieties; the remaining alkane parameters were directly transferred. However, when efforts were made to correct for the free energies of hydration, pair-specific LJ terms were introduced only for the hydroxyl O atoms. Changes were not made in the alkane LJ parameters, which were problematic, as stated above. This led to overcompensation in the case of the pair-specific LJ parameters for the interaction between the hydroxyl O atom and the water O atom. Accordingly, it is necessary to reconsider the implementation of pair-specific LJ parameters in the Drude polarizable force field.

If the pair-specific LJ approach is to be used to correct calculated hydration free energies within the CHARMM Drude polarizable force field, it is essential that these parameters be applied in a consistent way, which allows for the simultaneous representation of all classes of molecules. In addition, it would be useful for future force field developers if a general parametrization approach could be developed to allow for parameter optimization that is as systematic and straightforward as possible. With these goals in mind, the specific objectives of this work are as follows:

(1) The implementation of pair-specific LJ parameters in a hierarchical fashion, starting with the alkanes

(2) The development of a consistent set of pair-specific LJ parameters that give good reproduction of hydration free energies across all series of parametrized molecules

(3) The development of a reliable, systematic protocol for the determination of pair-specific LJ parameters.

## 2. Theory and Methods

The literature values of the hydration free energies calculated using the CHARMM Drude polarizable force field that are listed in Table 1 and illustrated in Figure 1 have been obtained from a series of distinct studies. To avoid any discrepancies introduced by small differences in free energy simulation methodologies and sampling, the first stage of this work was to recalculate the free energy of hydration for every molecule considered in this study using an identical protocol. Specifically, free energies of aqueous solvation were calculated *via* the free energy perturbation (FEP) method[59] using the staged protocol of Deng and Roux.[38] In this method, the LJ potential is separated into purely repulsive and attractive parts using the scheme originally developed by Weeks, Chandler, and Andersen (WCA).[60]

When a single solute molecule, $u$, is solvated in solvent $v$, with the coordinates of solute and solvent represented by $\mathbf{X}$ and $\mathbf{Y}$, respectively, the solute−solvent interaction potential, $E_{uv}(\mathbf{X},\mathbf{Y})$, comprises a short-range nonpolar contribution and a long-range electrostatic contribution:

$$E_{uv}(\mathbf{X},\mathbf{Y}) = E_{uv}^{\mathrm{np}}(\mathbf{X},\mathbf{Y}) + E_{uv}^{\mathrm{elec}}(\mathbf{X},\mathbf{Y}) \qquad (4)$$

The nonpolar contribution is given by the LJ equation (eq 1) and, using the WCA scheme, is separated into contributions due to the repulsive and attractive (dispersion) interactions, so that

$$E_{uv}^{\mathrm{np}}(\mathbf{X},\mathbf{Y}) = E_{uv}^{\mathrm{rep}}(\mathbf{X},\mathbf{Y}) + E_{uv}^{\mathrm{dis}}(\mathbf{X},\mathbf{Y}) \qquad (5)$$

Where the repulsive and attractive contributions to the LJ potential are given by eqs 6 and 7.

$$E_{ij}^{\mathrm{rep}}(r) = \begin{cases} \varepsilon_{ij}\left[\left(\frac{R_{\mathrm{min},ij}}{r}\right)^{12} - 2\left(\frac{R_{\mathrm{min},ij}}{r}\right)^{6} + 1\right] & r \geq R_{\mathrm{min},ij} \\ 0 & r \leq R_{\mathrm{min},ij} \end{cases} \qquad (6)$$

$$E_{ij}^{\mathrm{dis}}(r) = \begin{cases} \left[\left(\frac{R_{\mathrm{min},ij}}{r}\right)^{12} - 2\left(\frac{R_{\mathrm{min},ij}}{r}\right)^{6}\right] & r \geq R_{\mathrm{min},ij} \\ 0 & r \leq R_{\mathrm{min},ij} \end{cases} \qquad (7)$$

With the WCA scheme applied, the total potential energy of the system can be written as

$$E(\mathbf{X},\mathbf{Y}) = E_u(\mathbf{X}) + E_v(\mathbf{Y}) + E_{uv}^{\mathrm{elec}}(\mathbf{X},\mathbf{Y}) + E_{uv}^{\mathrm{rep}}(\mathbf{X},\mathbf{Y}) + E_{uv}^{\mathrm{dis}}(\mathbf{X},\mathbf{Y}) \qquad (8)$$

where $E_u$ is the internal potential energy of the solute molecule, $E_v$ is the solvent potential energy, and $E_{uv}$ represents the interaction between solvent and solute molecules, with the three terms corresponding to the Coulomb electrostatic, LJ-WCA core repulsion, and LJ-WCA dispersive attraction, respectively. For the free energy perturbation calculation, coupling between the initial and final states ($E_a$ and $E_b$) is achieved by means of a staging parameter. For both the electrostatic and dispersive interactions, a simple linear coupling of the initial and final states is used, with coupling parameters denoted $\lambda$ and $\xi$ (eqs 9 and 10).

$$E^{\mathrm{elec}}(\lambda) = (1-\lambda)E_a^{\mathrm{elec}} + \lambda E_b^{\mathrm{elec}} \qquad (9)$$

$$E^{\mathrm{dis}}(\xi) = (1-\xi)E_a^{\mathrm{dis}} + \xi E_b^{\mathrm{dis}} \qquad (10)$$

For the solute−solvent core repulsion term, such linear scaling is not practical, and the repulsion term is instead transformed into a soft-core potential using the nonlinear staging parameter, $s$:

$$E_{ij}^{\mathrm{rep}}(r,s) = \begin{cases} \left\{\frac{R_{\mathrm{min}}^{12}}{[r^2 + (1-s)^2R^2\mathrm{min}]^6} - 2\frac{R_{\mathrm{min}}^{6}}{[r^2 + (1-s)^2R^2\mathrm{min}]^3}\right\} & r \leq R_{\mathrm{min}}\sqrt{1-(1-s)^2} \\ 0 & r \geq R_{\mathrm{min}}\sqrt{1-(1-s)^2} \end{cases} \qquad (11)$$

With the formulation in place, the reversible work corresponding to the insertion of the fully interacting solute into the solvent is calculated in three steps using three distinct staging parameters $s$, $\xi$, and $\lambda$. Initially, the solute−solvent core repulsion is progressively introduced (eq 12), followed by the dispersion interaction (eq 13), and finally the electrostatic interaction (eq 14). The total solvation free energy is then the sum of these three terms.

$$\Delta G^{\mathrm{rep}} \equiv E(s=0,\xi=0,\lambda=0) \rightarrow E(s=1,\xi=0,\lambda=0) \qquad (12)$$

$$\Delta G^{\mathrm{dis}} \equiv E(s=1,\xi=0,\lambda=0) \rightarrow E(s=1,\xi=1,\lambda=0) \qquad (13)$$

$$\Delta G^{\mathrm{elec}} \equiv E(S=1,\xi=1,\lambda=0) \rightarrow E(s=1,\xi=1,\lambda=1) \qquad (14)$$

The computational details were identical to those described elsewhere,[30] but with the simulation time extended to 50 ps of equilibration and 100 ps of production for a given value of the coupling and/or staging parameter (with coordinates saved every 0.1 ps), and all free energy values presented as the average of five (rather than three) separate calculations.

A long-range correction[61] was also included to account for errors introduced by the truncation of LJ interactions. To calculate this long-range correction, for every calculated value of the hydration free energy, a single simulation of a single solute molecule in a box of 250 SWM4-NDP[25] water molecules was run for 50 ps of molecular dynamics in the NVT ensemble, during which coordinates were saved every 1 ps. Following completion of the MD simulation, coordinates were extracted from the final 30 ps of the CHARMM trajectory file, and energies were calculated for each set of coordinates using two different nonbonded interaction cutoff schemes. In the first scheme, nonbond pair lists were maintained to 14 Å with a cutoff of 12 Å used for both electrostatic and van der Waals (vdW) terms, with the latter truncated via an atom-based switch algorithm. In the second scheme, the only differences were that nonbond pair lists were maintained to 54 Å, and a cutoff of 50 Å was used. The difference in the vdW interaction energy calculated using the two nonbonded interaction cutoff schemes, averaged over all sets of coordinates, was taken as the long-range correction. The longer cutoff used in these calculations (50 Å) was significantly larger than that used in previous work, where nonbond pair lists were maintained to 36 Å and a cutoff of 32 Å was used.[30] The motivation for this change will be discussed in detail in the Results section. It should be noted that the box of 250 SWM4-NDP water molecules used in these calculations has a side length of approximately 20 Å. When a nonbond cutoff of 50 Å (or indeed 32 Å) is used, this means that periodic images of the solvent box must be used to calculate the total nonbond interactions. Each of these periodic images also includes one copy of the solute molecule, and so the total nonbond interaction energy includes a contribution due to solute−solute interactions. In practice, however, this contribution is small. The nearest solute image to the original solute molecule will be at a distance of 20 Å, and there will be six such images at this distance. Taking butane as an example, the solute-image solute interaction energy will be around −0.0005 kcal/mol per image, totaling −0.003 kcal/mol. Images at greater distances will have an even smaller impact. In addition, these solute molecules are occupying space that would otherwise be occupied by water molecules. A single butane molecule has a molecular volume of 160.5 Å,[26] which is equivalent to the volume occupied by 5.3 water molecules.[25] At a distance of 20 Å, 5.3 water molecules would contribute around −0.0003 kcal/mol to the total interaction energy. Overall, it can therefore be said that the overall error introduced by the presence of a single solute image at a distance of 20 Å is −0.0002 kcal/mol. Errors of this magnitude will have negligible impact on the final calculated results.

The computational method for calculation of the long-range correction described above has been applied in previous simulations involving the CHARMM Drude polarizable force field.[27,30,31,33] To evaluate the quality of this long-range correction calculation, the long-range correction has also been evaluated analytically[37,44,62] by solving eq 15.

$$E_{\mathrm{LRC}} = \sum_i 4\pi\rho\varepsilon \int_{r_{\mathrm{on}}}^{\infty} \left[\left(\frac{R_{\mathrm{min}}}{r}\right)^{12} - 2\left(\frac{R_{\mathrm{min}}}{r}\right)^{6}\right] S(r) r^2 \, \mathrm{d}r \tag{15}$$

where $i$ runs over all solute atoms, $r$ is the distance from solute atom $i$, $\rho$ is the number density of solvent molecules, $\varepsilon$ and $R_{\mathrm{min}}$ are the LJ parameters between atom $i$ and the O atom of the solvent water molecule (the H atoms of the SWM4-NDP water model have no LJ parameters), $S(r)$ is the switching function used to reduce smoothly the interaction from its full value to 0, and $r_{\mathrm{on}}$ is the distance at which the switching function is turned on. For this approach to be valid, it is required that the solvent radial distribution function $g(r) = 1$ at all points beyond $r_{\mathrm{on}}$. This is known to be true for the SWM4-NDP water model.[25]

The simulations described above were all performed using the program CHARMM[63] without the inclusion of any pair-specific LJ parameters. The same procedure was also used to calculate an initial, uncorrected, hydration free energy for any molecule that had not had its hydration free energy evaluated as part of a previous study.

**2.1. Pair-Specific LJ Parameter Determination.** Precise calculation of hydration free energies via the FEP method described above is a computationally intensive process, and it would be impractical to derive new pair-specific LJ parameters by scanning over ranges of $R_{\mathrm{min}}$ and $\varepsilon$ and using FEP to calculate the hydration free energy for every parameter combination. Instead, a method is implemented to provide an initial assessment of the approximate values of $R_{\mathrm{min}}$ and $\varepsilon$ that are likely to yield hydration free energies in good agreement with experimental results, so that the FEP calculation of actual hydration free energies can be reduced to only a small number of new parameter sets. To achieve this, initial molecular dynamics (MD) simulations were performed on each of the solute molecules in a box of 250 SWM4-NDP[25] water molecules for 150 ps at a temperature of 298 K in the NPT ensemble, with periodic boundary conditions (PBC) and the SHAKE algorithm[64] used to constrain covalent bonds to hydrogen. Electrostatic interactions were treated using particle-mesh Ewald (PME) summation[65] with a coupling parameter of 0.34 and a sixth order spline for mesh interpolation. All simulations used the standard CHARMM Drude polarizable force field parameters, as described in the respective publications,[26,27,30] and included no pair-specific LJ parameters. A time step of 1 fs was employed, and coordinates were saved to a trajectory file every 100 steps.

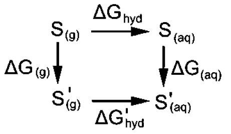


***Figure 2.*** Thermodynamic cycle for calculating the free energy of hydration with a perturbed set of LJ parameters, $\Delta G'_{\mathrm{hyd}}$, from the free energy of hydration with the original set of LJ parameters, $\Delta G_{\mathrm{hyd}}$. S indicates the solute represented using the original set of LJ parameters; S′ indicates the solute represented using the perturbed set of LJ parameters.

Once these MD simulations were complete, the free energy changes associated with changing the LJ parameters could be calculated. The LJ parameters used in the original MD simulation were first used to evaluate the solute−solvent interaction energy for every set of coordinates saved to the trajectory file. The LJ parameters used in the original MD simulation were then modified, the trajectory file was reread, and, for every set of coordinates, the solute−solvent interaction energy was re-evaluated using the new set of parameters. The difference in the solute−solvent interaction energies obtained using the original and modified LJ parameters was then used to estimate the free energy change associated with modifying the parameters. Once the free energy change for modifying the parameters in aqueous solution is obtained, it is straightforward to obtain the hydration free energy of the solute with the new LJ parameters by considering the thermodynamic cycle in Figure 2.

The free energy of hydration associated with the new set of LJ parameters, $\Delta G'_{\mathrm{hyd}}$, can be calculated from eq 16.

$$\Delta G'_{\mathrm{hyd}} = \Delta G_{\mathrm{hyd}} + \Delta G_{(\mathrm{aq})} - \Delta G_{(\mathrm{g})} \tag{16}$$

Because, by design, only the parameters affecting interactions between the solute and the solvent are modified, $\Delta G_{(\mathrm{g})} = 0$ such that the free energy change associated with modifying the parameters in aqueous solution, $\Delta G_{(\mathrm{aq})}$, is sufficient to provide for the difference between $\Delta G_{\mathrm{hyd}}$ and $\Delta G'_{\mathrm{hyd}}$. The method described above for the calculation of $\Delta G_{(\mathrm{aq})}$ is highly approximate because, in reality, the system will reorganize itself in response to any parameter change that changes the interaction energies and forces, whereas the approach outlined here assumes that the solvent structure around the solute is unaffected by the change in parameters. However, this technique is sufficient to provide a first approximation of parameter values that will yield a reasonable hydration free energy, and the impact of new parameter values can be assessed in a matter of seconds, rather than the approximately 2400 h of CPU time required to evaluate a single hydration free energy using the full method outlined above.

Once this approximate method had been used to identify a set of pair-specific LJ parameters appropriate for calculation of the hydration free energy for a given solute, its free energy of hydration was evaluated using the full FEP method described above. Three independent FEP calculations were performed, and the resulting hydration free energy values were averaged to give a final result. This result was then compared to the relevant experimental value. If necessary, the pair-specific LJ parameters were adjusted again and the hydration free energy re-evaluated, with this process repeated until satisfactory agreement with the experiment was obtained. During parametrization of the CHARMM Drude polarizable force field, the aim is generally that final calculated values should be within ∼2% of the corresponding experimental values. In this work, where experimental target values can be extremely small, and uncertainties in calculated values relatively large, such an approach is less reasonable. Cyclopentane, for example, has an experimental hydration free energy of 1.20 kcal/mol: a 2% target would require a calculated value to be between 1.18 kcal/mol and 1.22 kcal.mol. Given that the uncertainty in the calculated value of $\Delta G_{\text{hyd}}$ for cyclopentane with no pair-specific LJ parameters is 0.05 kcal/mol (Table 1), this level of accuracy is unrealistic. Rather, a goal where the final calculated hydration free energies should be within 0.1 kcal/mol of the corresponding experimental value is more reasonable. Once satisfactory agreement with the experiment had been obtained, further FEP calculations were performed so that the final hydration free energy values presented in this work are the average of five individual calculations. The error in each calculation is given as the standard deviation of the mean calculated over 500 iterations of a bootstrap procedure using software by Wessa.[66]

To evaluate the effect that the introduction of pair-specific LJ parameters would have on other calculated properties, solute−water heterodimeric complexes were examined. The methods used and results obtained are described in the Supporting Information that accompanies this paper.

**2.2. Testing the Need for Pair-Specific LJ Parameters.** It has been shown in the past that the use of pair-specific LJ parameters allows for the correction of hydration free energies when LJ parameters derived to reproduce liquid phase thermodynamic properties are unable to, and this study aims to exploit this fact. There is, however, an important question that must also be addressed during this work: are pair-specific LJ parameters really essential or, as some have suggested, would it be possible, by including $\Delta G_{\text{hyd}}$ values as target data in the initial parameter optimization, to find a set of LJ parameters that are able to reproduce accurately both the liquid phase thermodynamic data and solvation free energies simultaneously?

In an attempt to answer this question, the final pair-specific LJ parameters developed in this study were broken down into their constituent parts using the inverse of the standard LJ combining rules:

$$\frac{R_{\text{min}}}{2}, i = R_{\text{min}} - \frac{R_{\text{min}}}{2}, \text{ODW} \tag{17}$$

$$\varepsilon_i = -\frac{\varepsilon^2}{\varepsilon_{\text{ODW}}} \tag{18}$$

where $R_{\text{min}}$ and $\varepsilon$ are the pair-specific LJ parameter values and the ODW atom LJ parameters are fixed, thereby transferring the whole of the effect of the pair-specific LJ parameters onto the solute heavy atoms. In this way, it was possible to generate a new set of atomic LJ parameters, $R_{\text{min}}/2$ and $\varepsilon_i$, for every atom type considered in this study. Once this had been done, a series of calculations were performed to evaluate the molecular volume ($V_{\text{m}}$) and enthalpy of vaporization ($\Delta H_{\text{vap}}$) of each of four alkane and five ether molecules, to assess whether these new pair-specific LJ parameters would be appropriate for use in both the bulk liquid and aqueous solution, indicating that one set of parameters would be sufficient in both cases, and that specific heavy atom−ODW LJ parameters would be unnecessary. [When eq 18 is applied for the calculation of energies or forces (e.g., as in eq 1), $\varepsilon$ has a positive value. However, within the CHARMM parameter file, by convention $\varepsilon$ is always shown as negative, in both the NONBOND and NBFIX sections. For the sake of convenience, the CHARMM parameter file notation is used throughout this paper, and $\varepsilon$ values are always shown to be negative.] To calculate $V_{\text{m}}$ and $\Delta H_{\text{vap}}$ for each molecule, 10 liquid phase molecular dynamics simulations of 150 ps duration were performed. All 10 liquid phase simulations were commenced from an identical pre-equilibrated box of 128 molecules, with a random number seed used to assign different initial velocities in each case. The first 50 ps were treated as equilibration, with the remaining 100 ps used for analysis. Volumes and energies were averaged over all 10 simulations, and the gas phase contribution to the heat of vaporization was calculated from a single simulation of 2.5 ns, with 0.5 ns used for equilibration and 2.0 ns for analysis. All simulations were performed at the temperatures reported in Table 2.

## 3. Results

**3.1. The Long Range Correction.** As noted above, in previous studies where the CHARMM Drude polarizable force field has been used to calculate hydration free energies, a cutoff of 32 Å has been used in the evaluation of the long-range correction associated with the truncation of the LJ interactions. In this study, the effect of the cutoff on the total long-range correction was examined, and the results can be seen in Figure 3, where long-range corrections have been calculated for progressively larger molecules. While using a cutoff of 32 Å (denoted by the vertical line in Figure 3) captures the majority of the long-range correction, it is clear that at 32 Å the long-range correction has not yet reached convergence. To achieve convergence (to two decimal places) for all of the molecules considered in this study, it was necessary to use a cutoff of at least 50 Å. The final long-range correction values obtained for all molecules in this study, both with and without pair-specific LJ parameters, are presented in Table 3 along with long-range correction values calculated analytically. The analytically calculated values can be considered the “correct” values, and it is encouraging to note that the numerically calculated values are very close to the analytically calculated values, with an average error of −0.011 kcal/mol and a maximum error of −0.018 kcal/mol. Such small errors will have minimal impact on the final hydration free energies, and it can be concluded that the numerical method is valid for the evaluation of the long-range correction.

***Table 2.*** $V_m$ and $\Delta H_{vap}$ Calculated Using LJ Parameters Obtained from the Pair-Specific LJ Parameters Calculated in This Work, and Compared to $V_m$ and $\Delta H_{vap}$ Calculated Using the Standard CHARMM Drude Polarizable Force Field LJ Parameters

| | *T*/K | experimental | standard LJ | % err | pair-specific LJ | % err |
|---|---|---|---|---|---|---|
| | | | Molecular Volumes | | | |
| ETHA | 184.55 | 91.8 | 91.6 ± 0.3 | −0.2 | 95.6 ± 1.7 | 4.1 |
| PROP | 231.10 | 125.7 | 124.5 ± 0.4 | −1.0 | 136.7 ± 1.8 | 8.8 |
| BUTA | 272.65 | 160.5 | 160.9 ± 0.3 | 0.2 | 182.8 ± 1.8 | 13.9 |
| IBUT | 261.43 | 162.5 | 160.6 ± 0.3 | −1.2 | 187.4 ± 3.0 | 15.3 |
| THF | 298.15 | 135.6 | 134.8 ± 0.4 | −0.6 | 148.4 ± 1.6 | 9.5 |
| THP | 298.15 | 162.3 | 163.8 ± 0.8 | 0.9 | 188.7 ± 1.8 | 16.3 |
| DMOE | 298.15 | 173.6 | 178.1 ± 0.9 | 2.6 | 194.3 ± 1.3 | 11.9 |
| DME | 248.34 | 104.9 | 104.2 ± 0.8 | −0.7 | 108.3 ± 1.0 | 3.2 |
| MEET | 273.20 | 137.5 | 140.2 ± 0.8 | 2.0 | 152.8 ± 1.4 | 11.1 |
| | | | Heats of Vaporization | | | |
| ETHA | 184.55 | 3.53 | 3.42 ± 0.01 | −3.1 | 3.23 ± 0.03 | −8.5 |
| PROP | 231.10 | 4.51 | 4.48 ± 0.01 | −0.7 | 3.67 ± 0.02 | −18.6 |
| BUTA | 272.65 | 5.37 | 5.41 ± 0.03 | 0.7 | 3.66 ± 0.02 | −31.8 |
| IBUT | 261.42 | 5.12 | 5.03 ± 0.02 | −1.8 | 3.71 ± 0.04 | −27.5 |
| THF | 298.15 | 7.65 | 7.69 ± 0.03 | 0.9 | 5.66 ± 0.04 | −26.0 |
| THP | 298.15 | 8.26 | 8.41 ± 0.04 | 1.8 | 5.59 ± 0.04 | −32.3 |
| DMOE | 298.15 | 8.79 | 8.67 ± 0.07 | −1.4 | 6.82 ± 0.04 | −22.4 |
| DME | 248.34 | 5.14 | 5.18 ± 0.02 | 0.8 | 4.51 ± 0.02 | −12.3 |
| MEET | 280.60 | 5.90 | 5.85 ± 0.04 | −0.8 | 4.68 ± 0.04 | −20.7 |

**3.2. Parametrization Strategy.** One of the key objectives of this work was to obtain not only a set of useable parameters but also a reliable method by which they should be obtained. The initial strategy employed was to vary $R_{min}$ until good agreement was obtained between the calculated and experimental hydration free energies. In particular, since all but one of the calculated hydration free energies were more favorable than their experimental equivalents, it was anticipated that increasing $R_{min}$ would be a good general strategy for making calculated free energies less favorable. For polar molecules, this was based on the assumption that, by increasing the radius at which the most favorable interaction occurs, atom pairs having favorable electrostatic interactions (specifically, hydrogen bonding interactions involving water molecules) would be pushed further apart, and these favorable electrostatic interactions would decrease. However, in the case of the nonpolar alkanes, such an approach is not appropriate because the LJ term dominates the free energy of aqueous solvation. For example, in the acyclic alkanes, an increase in $R_{min}$ resulted in a more favorable free energy of hydration, as shown for butane in Figure 4.

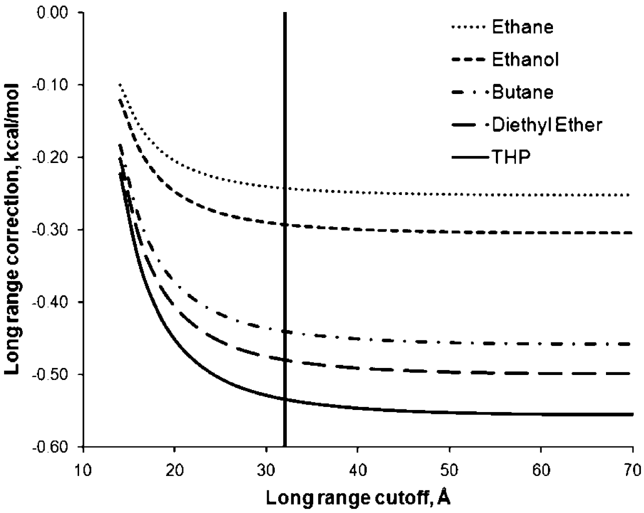


***Figure 3.*** Dependence of the long-range LJ correction on the magnitude of the cutoff used. The vertical line indicates a cutoff of 32 Å, the previous "standard value" used in calculating the long-range correction with the CHARMM Drude polarizable force field.

This effect can be explained by considering the functional form of the LJ term (eq 1): Figure 5 shows two such LJ curves in which $R_{min}$ differs, but $\varepsilon$ is unchanged. Comparison of these two curves shows that an atom−atom pair with a separation, $r$, greater than $r_{int}$, the point at which the two curves intersect, will have a more favorable LJ interaction energy when $R_{min} = R_{min2}$ than when $R_{min} = R_{min1}$. An atom pair with a separation, $r$, less than $r_{int}$, will have a less favorable interaction when $R_{min} = R_{min2}$ than when $R_{min} = R_{min1}$. Given the large number of atom−atom pairs with distances greater than $r_{int}$, an increase in $R_{min}$ from $R_{min1}$ to $R_{min2}$ usually results in a more favorable total interaction. This in turn leads to the more favorable free energy of solvation of the alkanes with larger $R_{min}$ values on the C atoms, because the solvation free energy has a significant contribution from the LJ term as compared to more polar molecules. It is not until $R_{min}$ become so large that it causes significant short-range atom−atom repulsion that the LJ energy starts to become less favorable. Alternatively, increasing $\varepsilon$ without changing $R_{min}$ (Figure 5) yields the more intuitive result where the overall LJ surface is more favorable at all atom−atom distances with the LJ interaction energy >0. Importantly, varying $\varepsilon$ also does not significantly impact the repulsive wall, which in the present study was that obtained from parameters based on the pure solvent or crystal simulations.

With these observations in mind a modified parametrization strategy was developed, having three distinct stages.

(1) For polar molecules, an attempt is made to correct the hydration free energy by varying only $R_{min}$ of

***Table 3.*** Calculated Long Range Corrections, in kcal/mol, for Molecules Considered in This Work

| | numerically calculated long range correction[a] | | analytically calculated long range correction | |
|---|---|---|---|---|
| molecule | without pair-specific LJ parameters | with pair-specific LJ parameters | without pair-specific LJ parameters | with pair-specific LJ parameters |
| | | Alkanes | | |
| CPEN | −0.505 ± 0.002 | −0.456 ± 0.002 | −0.519 | −0.468 |
| CHEX | −0.617 ± 0.002 | −0.575 ± 0.002 | −0.634 | −0.592 |
| ETHA | −0.250 ± 0.001 | −0.243 ± 0.001 | −0.255 | −0.248 |
| PROP | −0.353 ± 0.001 | −0.328 ± 0.002 | −0.361 | −0.334 |
| BUTA | −0.456 ± 0.002 | −0.409 ± 0.002 | −0.467 | −0.421 |
| IBUT | −0.441 ± 0.001 | −0.391 ± 0.002 | −0.455 | −0.405 |
| NEOP | −0.549 ± 0.001 | −0.487 ± 0.001 | −0.568 | −0.505 |
| | | Alcohols | | |
| MEOH | −0.225 ± 0.001 | −0.225 ± 0.001 | −0.229 | −0.229 |
| ETOH | −0.303 ± 0.001 | −0.280 ± 0.002 | −0.311 | −0.288 |
| PRO2 | −0.392 ± 0.001 | −0.347 ± 0.001 | −0.404 | −0.357 |
| BUO2 | −0.494 ± 0.002 | −0.430 ± 0.001 | −0.511 | −0.444 |
| PRO1 | −0.402 ± 0.002 | −0.357 ± 0.001 | −0.414 | −0.368 |
| BUO1 | −0.504 ± 0.002 | −0.440 ± 0.001 | −0.521 | −0.454 |
| | | Ethers | | |
| THF | −0.455 ± 0.002 | −0.408 ± 0.002 | −0.464 | −0.415 |
| THP | −0.553 ± 0.002 | −0.475 ± 0.001 | −0.564 | −0.484 |
| DEE | −0.495 ± 0.001 | −0.448 ± 0.003 | −0.506 | −0.458 |
| DMOE | −0.550 ± 0.001 | −0.499 ± 0.001 | −0.567 | −0.512 |
| DME | −0.309 ± 0.001 | −0.296 ± 0.002 | −0.317 | −0.303 |
| MEE | −0.401 ± 0.001 | −0.371 ± 0.002 | −0.411 | −0.381 |

[a] Calculated values averaged over five independent simulations, with errors as ± 1 standard deviation.

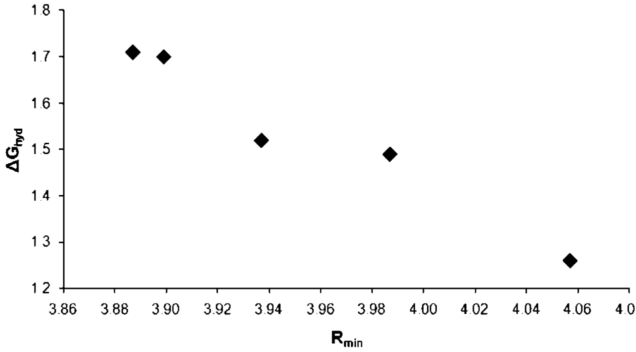


***Figure 4.*** Calculated hydration free energy of butane as a function of $R_{\mathrm{min}}$ for the CD32A-ODW pair, with all other LJ parameters fixed. $R_{\mathrm{min}}$ in Å, $\Delta G_{\mathrm{hyd}}$ in kcal/mol.

heavy atom−ODW pairs, up to a maximum $\Delta R_{\mathrm{min}}$ of 0.1 Å: if the calculated $\Delta G_{\mathrm{hyd}}$ in the absence of pair-specific LJ parameters is too favorable, only increasing $R_{\mathrm{min}}$ is considered; if the calculated $\Delta G_{\mathrm{hyd}}$ in the absence of pair-specific LJ parameters is not favorable enough, only decreasing $R_{\mathrm{min}}$ is considered.

(2) In the case of nonpolar molecules, an attempt is made to correct the free energy of hydration by varying only $\varepsilon$ of heavy atom−ODW pairs.

(3) If either 1 or 2 is unsuccessful, an attempt is made to correct the hydration free energy by increasing both $R_{\mathrm{min}}$ and $\varepsilon$ of heavy atom−ODW atom pairs simultaneously.

To date, such an approach has been sufficient to give pair-specific LJ parameters that provide good agreement with experimental data in every case, with one exception. It is anticipated that, in the future, in the small number of cases where this scheme will not be successful, the molecules in question will need to be approached on a case-by-case basis: the only molecule for which pair-specific LJ parameters could not be obtained using this scheme in the present work will be discussed in detail below. All pair-specific LJ parameters obtained in this work are listed in Table 4.

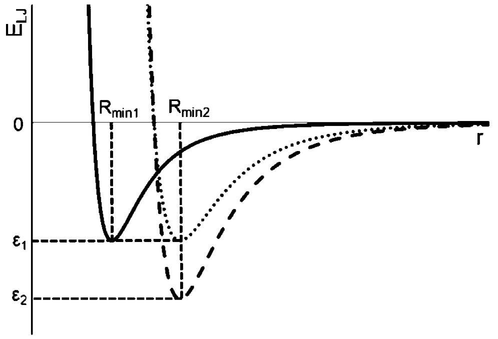


***Figure 5.*** Example LJ interaction energy curves. Comparing the two curves with $\varepsilon = \varepsilon_1$: if the two curves intersect at a point $r_{\mathrm{int}}$, then all interactions with $r > r_{\mathrm{int}}$ will become more favorable on moving from $R_{\mathrm{min1}}$ to $R_{\mathrm{min2}}$; all interactions with $r < r_{\mathrm{int}}$ will become less favorable on moving from $R_{\mathrm{min1}}$ to $R_{\mathrm{min2}}$. Comparing the two curves with $R_{\mathrm{min}} = R_{\mathrm{min2}}$: moving from $\varepsilon_1$ to $\varepsilon_2$ results in interactions becoming more favorable at all values of $r$.

**3.3. Hydration Free Energies.** A total of 19 molecules were chosen to comprise the "parametrization set" (Figure 6), the set of molecules that would be used to develop the pair-specific LJ parameters. With the aim of creating a consistent, systematic set of pair-specific LJ parameters for use across all molecules, it was necessary to take the alkanes as a starting point. For the alkanes, seven molecules were considered as part of the parametrization process: the acyclic alkanes ETHA, PROP, BUTA, IBUT, and NEOP and the cyclic alkanes CPEN and CHEX. The first step of the parametrization involved the development of pair-specific LJ parameters for the ethane methyl C atoms (Ca, Figure 6). Once these parameters had been developed, they were

***Table 4.*** Final Pair-Specific LJ Parameters, and Comparison to LJ Parameters Obtained Using Standard Combining Rules[a]

| atom | atom | atom | standard LJ parameters | | pair-specific LJ Parameters | | change in LJ parameters | |
|---|---|---|---|---|---|---|---|---|
| name | type 1 | type 2 | $\varepsilon$ | $R_{min}$ | $\varepsilon$ | $R_{min}$ | $\Delta\varepsilon$ | $\Delta R_{min}$ |
| Ca | CD33A | ODW | −0.1283 | 3.8269 | −0.1233 | 3.8269 | 0.0050 | 0.0000 |
| Cb | CD32A | ODW | −0.1087 | 3.8869 | −0.0817 | 3.8869 | 0.0270 | 0.0000 |
| *Cc* | CD31A | ODW | −0.0681 | 3.9869 | −0.0211 | 3.9869 | 0.0470 | 0.0000 |
| Cd | CD30A | ODW | −0.0650 | 3.9869 | −0.0050 | 4.1869 | 0.0600 | 0.2000 |
| Ce | CD325A | ODW | −0.1125 | 3.8069 | −0.0965 | 3.8069 | 0.0160 | 0.0000 |
| Cf | CD326A | ODW | −0.1087 | 3.8869 | −0.0992 | 3.8869 | 0.0095 | 0.0000 |
| Ch | CD33E | ODW | −0.1481 | 3.7869 | −0.1431 | 3.7869 | 0.0050 | 0.0000 |
| Ci | CD32E | ODW | −0.1067 | 3.8069 | −0.0797 | 3.8069 | 0.0270 | 0.0000 |
| Cj | CD325B | ODW | −0.1125 | 3.8069 | −0.0925 | 3.8069 | 0.0200 | 0.0000 |
| Ck | CD326B | ODW | −0.1087 | 3.7969 | −0.0827 | 3.7969 | 0.0260 | 0.0000 |
| Ob | OD31B | ODW | −0.1779 | 3.5269 | −0.1779 | 3.4969 | 0.0000 | −0.0300 |
| Oc | OD30A | ODW | −0.1125 | 3.5269 | −0.0919 | 3.5469 | 0.0206 | 0.0200 |
| Od | OD305A | ODW | −0.1299 | 3.5069 | −0.1299 | 3.5269 | 0.0000 | 0.0200 |
| Oe | OD306A | ODW | −0.1299 | 3.5269 | −0.1299 | 3.5469 | 0.0000 | 0.0200 |
| N/A | CD315A | ODW | −0.0822 | 3.7869 | −0.0662 | 3.7869 | 0.0160 | 0.0000 |
| N/A | CD315B | ODW | −0.0822 | 3.7869 | −0.0622 | 3.7869 | 0.0200 | 0.0000 |
| N/A | CD316A | ODW | −0.0822 | 3.7869 | −0.0727 | 3.7869 | 0.0095 | 0.0000 |

[a] $\varepsilon$ in kcal/mol, $R_{min}$ in Å. Atom names are as listed in Figure 6: atom types CD315A, CD315B, and CD316A are from the test set molecules CPNM, TF2M, and CHXM, respectively. No pair-specific LJ parameters were required for atoms Cg or Oa.

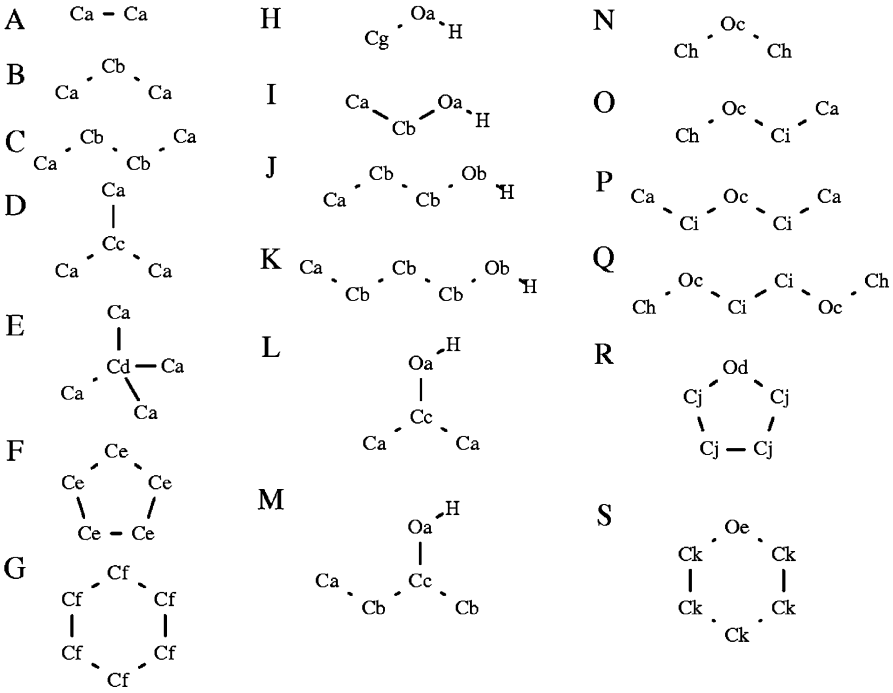


***Figure 6.*** Compounds used in development of pair-specific LJ parameters: (a) ethane, ETHA; (b) propane, PROP; (c) butane, BUTA; (d) isobutane, IBUT; (e) neopentane, NEOP; (f) cyclopentane, CPEN; (g) cyclohexane, CHEX; (h) methanol, MEOH; (i) ethanol, ETOH; (j) propan-1-ol, PRO1; (k) butan-1-ol, BUO1; (l) propan-2-ol, PRO2; (m) butan-2-ol, BUO2; (n) dimethyl ether, DME; (o) methyl ethyl ether, MEET; (p) diethyl ether, DEET; (q) 1,2-dimethoxyethane, DMOE; (r) tetrahydrofuran, THF; (s) tetrahydropyran, THP.

then used in the development of parameters for the Cb atoms, based on propane and butane; the Cc atom, based on isobutane; and the Cd atom, based on neopentane. While the C atom in CPEN was always treated as having a different atom type from the acyclic $CH_2$ C atoms, CHEX C atoms were initially assigned the Ca atom type. However, it was not possible to obtain a set of pair-specific LJ parameters that gave good agreement across both the acyclic alkanes and CHEX, and ultimately, the C atoms of CHEX were assigned their own atom type. In this way, it was possible to construct a consistent set of parameters that gave good agreement with experimental $\Delta G_{hyd}$ values across the whole range of alkane molecules considered as part of the parametrization process (Table 1). Overall, the average error in the calculated hydration free energy has been reduced from −0.91 to −0.05 kcal/mol, with the root-mean-square deviation (rmsd) reduced from 1.02 to 0.10 kcal/mol, indicating that the systematically too-favorable prediction of alkane hydration free energies has been corrected. In general, the agreement with experimental results obtained using the new pair-specific LJ parameters is excellent across all alkane molecules, with only NEOP (with a deviation of −0.25 kcal/

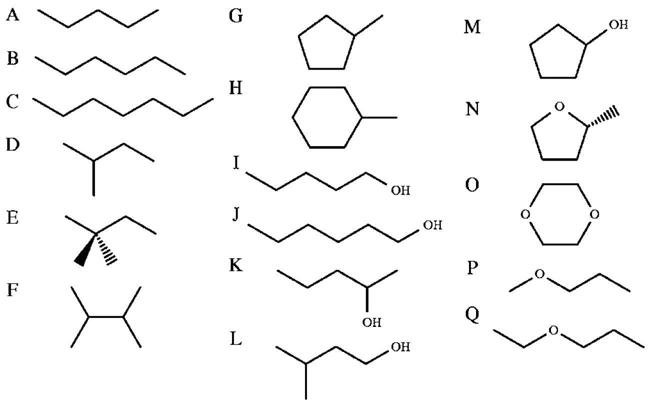


*Figure 7.* Compounds used for testing pair-specific LJ parameters: (a) pentane, PENT; (b) hexane, HEXA; (c) heptane, HEPT; (d) 2-methylbutane, BU2M; (e) 2,2-dimethylbutane, BU22M; (f) 2,3-dimethylbutane, BU23M; (g) methylcyclopentane, CPNM; (h) methylcyclohexane, CHXM; (i) pentan-1-ol, PEO1; (j) hexan-1-ol, HXO1; (k) pentan-2-ol, PEO2; (l) 3-methylbutan-1-ol, B3MO1; (m) cyclopentanol, CPOH; (n) 2-(R)-methyl tetrahydrofuran, MTHF; (o) 1,4-dioxane, DIOX; (p) methyl propyl ether, MPET; (q) ethyl propyl ether, EPET.

mol from the experimental value) giving a deviation with magnitude greater than 0.07 kcal/mol from the corresponding experimental value. Moreover, the inclusion of pair-specific LJ parameters results in an accurate reproduction of the ordering of $\Delta G_{hyd}$ values. The LJ parameters obtained using the standard combining rules incorrectly predicted that $\Delta G_{hyd}$ values decrease with increasing chain length. When pair-specific LJ parameters are included, hydration free energies become less favorable with increasing chain length, in agreement with experimental results.

Examination of Table 4 reveals that the central C atom of NEOP (Cd in Figure 7; CHARMM atom type CD30A) is also the only alkane atom type for which it was necessary to break the "rules" for pair-specific LJ parameter development outlined above. The final pair-specific LJ parameters for Cd have $\Delta\varepsilon = 0.0600$ and $\Delta R_{min} = 0.2000$: for comparison, the largest change in any of the other alkane atom types is found in CD31A from IBUT (Cc, Figure 7), where $\Delta\varepsilon = 0.0470$ and $\Delta R_{min} = 0.0000$. Put simply, it appears that the CD30A atom of NEOP is being asked to do too much work. Before any pair-specific LJ parameters are added, NEOP gives the hydration free energy in worst agreement with experimental data (Table 1). In addition, the changes made to the methyl C atom (Ca) are extremely small, meaning that only the pair-specific LJ parameters for the CD30A atom type could be optimized to correct the calculated $\Delta G_{hyd}$. With this atom surrounded by methyl groups in NEOP, it is a significant distance from the nearest water molecules, thereby reducing the impact of any changes in the LJ parameters on $\Delta G_{hyd}$. While the magnitude of the difference upon moving from the combining rule to pair-specific LJ parameters is not ideal, the CD30A atom type does not appear in biomolecular systems, which are the ultimate target of this small molecule work, and so was not a great cause for concern.

It should be noted that two papers focused on the development of computational methods for estimating hydration free energies have reported experimental values of the hydration free energy for neopentane that are significantly different. Michielan et al. reported a value of 2.69 kcal/mol,[67] while Ooi et al. reported a value of 2.50 kcal/mol.[68] While Michielan et al. give no information on the source of the experimental value used in their work, Ooi et al. provide references to the original sources of their experimental data.[69,70] For this reason, the experimental hydration free energy of neopentane used in this work is that obtained from the work of Ooi et al.

The alkane parameters were then applied to the alcohol and ether molecules, with the logic being that pair-specific LJ parameters for atom types not included in the alkanes should be built on top of the alkane pair-specific LJ parameters, so as to yield a set of parameters that is consistent across all molecules.

For the alcohols, inclusion of the alkane pair-specific LJ parameters has a dramatic effect on the calculated hydration free energies (Table 1). For MEOH, ETOH, PRO2, and BUO2, which share an atom type for the hydroxyl O, no further pair-specific LJ parameters were required to yield an acceptable improvement in the calculated $\Delta G_{hyd}$ values. For the long chain primary alcohols PRO1 and BUO1, which possess a different O atom type than the other alcohols, the addition of the alkane pair-specific LJ parameters results in a slight overcorrection, making the $\Delta G_{hyd}$ values, which were initially too favorable, not favorable enough. Pair-specific LJ parameters were applied to the O atom to rectify this overcorrection (Table 4). The resulting set of pair-specific LJ parameters gave an average error for the alcohols of −0.06 kcal/mol and an rmsd of 0.32 kcal/mol, compared to an average error of −0.54 kcal/mol and an rmsd of 0.65 kcal/mol for the values obtained using the LJ parameters obtained from the standard combining rules.

For the ethers, the situation was complicated by the presence of several C atom types that do not appear in the alkanes, corresponding to the C atoms adjacent to the ether O atoms in the linear ethers. For these atom types, the *change* in the LJ parameters needed to obtain pair-specific LJ parameters for the corresponding alkane atom was retained for use in the ether atom types, resulting in pair-specific LJ parameters that differ in magnitude but show the same change relative to the combining rule LJ parameters. With these C atom pair-specific LJ parameters in place, it was a matter of adjusting only the Oc atom type pair-specific LJ parameters until optimal agreement with the experiment was obtained. For the cyclic ethers THF and THP, a similar approach was attempted, in which the *change* in LJ parameters for the C atoms was transferred directly from the corresponding atom types in CPEN and CHEX. Using such an approach, however, very large changes were required in the Od/Oe-ODW LJ parameters to obtain acceptable hydration free energies. These changes not only violated the rules outlined above for the derivation of pair-specific LJ parameters but also resulted in a significant worsening of the calculated gas phase heterodimer interactions with water molecules (Table S3 of the Supporting Information). Accordingly, for THF and THP, this approach was abandoned, and pair-specific LJ parameters for both the C and O atoms of both molecules were allowed to vary. The final set of pair-specific LJ parameters gave hydration free energies as shown in Table 1: the average error in the values calculated using the new pair-specific LJ parameters was 0.01 kcal/mol with an rmsd of 0.17 kcal/mol, compared to an average error of −0.95 kcal/mol and an rmsd of 1.21 kcal/mol in the values calculated without pair-specific LJ parameters.

Across all 19 molecules considered in the parametrization process, the average error in the $\Delta G_{\mathrm{hyd}}$ values calculated using the pair-specific LJ parameters is −0.03 kcal/mol, with an rmsd of 0.21 kcal/mol. For $\Delta G_{\mathrm{hyd}}$ values calculated without the inclusion of any pair-specific LJ parameters, the average error is −0.84 kcal/mol and the rmsd is 0.99 kcal/mol. Performing a Student's *t* test[71] results in the rejection of the null hypothesis that these two mean errors are the same ($P$ value $\leq 0.0001$): the difference between the average errors is statistically significant. Clearly, through the inclusion of pair-specific LJ parameters, the systematic error in the calculated $\Delta G_{\mathrm{hyd}}$ values has been eliminated, while at the same time the absolute error in the $\Delta G_{\mathrm{hyd}}$ values has also decreased.

To further ensure the utility of the pair-specific LJ parameters, the issue of sampling was considered: if free energy values are to be calculated accurately, it is important that all accessible conformations of a molecule and its aqueous environment be sampled to yield an adequate precision.[37] While torsional modes tend to be most problematic when it comes to achieving adequate sampling, even nontorsional relaxation times are on the order of 2−10 ps. With, in this case, 100 ps of sampling per coupling value, this results in 10−100 independent samples. To assess whether the use of 100 ps/window in the free energy calculations represents a sufficient level of sampling, FEP calculations were performed for ETOH and THF using the method described above with 500 ps rather than 100 ps of production MD for every value of the coupling and/or staging parameter. These calculations were performed using the final values of the pair-specific LJ parameters obtained in this work. For ETOH, the mean hydration free energy obtained over five independent calculations with the longer calculations was −4.73 ± 0.03 kcal/mol. The equivalent value obtained from the original, shorter, calculations was −4.81 ± 0.05 kcal/mol. Performing a Student's *t* test[71] with a significance level of 0.05 leads to the acceptance of the null hypothesis that the two means are the same ($P$ value = 0.234). The same conclusion is also reached for THF ($P$ value = 0.555) where the shorter simulations gave $\Delta G_{\mathrm{hyd}} = -3.58 \pm 0.07$ kcal/mol and the longer simulations gave $\Delta G_{\mathrm{hyd}} = -3.62 \pm 0.03$ kcal/mol. Overall, it can be concluded that, for these molecules, performing longer MD simulations has no statistically significant effect on the calculated hydration free energies and that the level of sampling used in the original calculations is adequate.

**3.4. Test Compounds.** To test the transferability of the parameters obtained above, simulations were performed on another 17 compounds (Figure 7): six acyclic alkanes, three linear (PENT, HEXA, HEPT) and three branched (BU2M, BU22M, BU23M); two cyclic alkanes (CPNM, CHXM); four acyclic alcohols, three linear (PEO1, PEO2, HXO1) and one branched (B3MO1); one cyclic alcohol (CPOH); two acyclic ethers (MPET, EPET); and two cyclic ethers (MTHF, DIOX). This test set was designed to include at least one example of every atom type for which pair-specific LJ parameters had been developed above. In total, 18 different atom types are represented within the test set. Fifteen of these were considered during the pair-specific LJ parameter optimization, with the remaining three having no pair-specific LJ parameters. For all 17 molecules, simulations were performed both with and without the pair-specific LJ parameters developed above. For the 15 atom types for which pair-specific LJ parameters had been explicitly parametrized, all of the pair-specific LJ parameters used in the simulation of these molecules were taken directly from Table 4. The three atom types for which pair-specific LJ parameters had not been explicitly calculated were the CHARMM atom types CD315B, CD315A, and CD316A, corresponding to the ring C atoms bonded to the substituent methyl groups in MTHF, CPNM (and CPOH), and CHXM, respectively. These atom types have LJ parameters that differ from other C atoms in their respective rings, which have the same atom types as the THF, CPEN, and CHEX ring C atoms.[30] In such cases, where pair-specific LJ parameters have not been optimized, pair-specific LJ parameters were introduced on the basis of the assumption that the *change* in the LJ parameters will be the same as the *change* needed to obtain pair-specific LJ parameters for the parent ring C atoms. Obtaining parameters by analogy in this manner is not a recommended procedure and generally yields suboptimal results. In this case, however, such an approach was deemed necessary to retain an objective test set. If the pair-specific LJ parameters for atom types present in the test set had been optimized, then the molecules containing these atoms types could no longer have been considered as part of the test set. It is anticipated that in future work where new pair-specific LJ parameters are required, such parameters would be obtained using the full optimization method outlined above. All parameters other than pair-specific LJ parameters had the standard CHARMM Drude polarizable force field values for alkanes, alcohols, and ethers.[26,27,30] A small number of dihedral and angle parameters that did not already exist within the CHARMM Drude polarizable force field were obtained by analogy to existing force field parameters. Again, such an approach is unlikely to yield high quality parameters but was deemed sufficient for the current test.

With the parameters in place, for each molecule, five independent calculations were performed to evaluate $\Delta G_{\mathrm{hyd}}$ using the FEP method described above. The final, average, value of $\Delta G_{\mathrm{hyd}}$ was then compared to the relevant experimental value, with a good reproduction of the experimental value taken to signify that the parameters are broadly transferable across a range of molecules.

The results of the calculations of hydration free energies on the test compounds are shown in Table 5. In all cases, the inclusion of the pair-specific LJ parameters results in a significant improvement in the calculated $\Delta G_{\mathrm{hyd}}$, with the largest error being −0.65 kcal/mol for both MTHF and

***Table 5.*** Free Energies of Hydration of Test Set Molecules

| molecule | experimental $\Delta G_{\mathrm{hyd}}$ | without pair-specific LJ parameters $\Delta G_{\mathrm{hyd}}$ | error | with pair-specific LJ parameters $\Delta G_{\mathrm{hyd}}$ | error |
|---|---|---|---|---|---|
| | | Alkanes | | | |
| PENT | 2.36[a] | 1.24 ± 0.09 | −1.12 | 2.61 ± 0.08 | 0.25 |
| HEXA | 2.48[a] | 0.85 ± 0.12 | −1.63 | 2.39 ± 0.12 | −0.09 |
| HEPT | 2.62[a] | 0.34 ± 0.10 | −2.28 | 2.81 ± 0.08 | 0.19 |
| BU2M | 2.38[b] | 0.55 ± 0.09 | −1.82 | 2.24 ± 0.05 | −0.14 |
| BU22M | 2.51[b] | 0.53 ± 0.15 | −1.98 | 1.95 ± 0.14 | −0.56 |
| BU23M | 2.34[b] | 0.87 ± 0.22 | −1.47 | 2.69 ± 0.12 | 0.36 |
| CPNM | 1.59[b] | 0.34 ± 0.07 | −1.25 | 1.64 ± 0.12 | 0.05 |
| CHXM | 1.70[b] | 0.31 ± 0.15 | −1.39 | 1.17 ± 0.08 | −0.53 |
| | | Alcohols | | | |
| PEO1 | −4.57[b] | −5.73 ± 0.07 | −1.16 | −4.66 ± 0.06 | −0.09 |
| HXO1 | −4.40[b] | −5.79 ± 0.25 | −1.39 | −4.81 ± 0.14 | −0.41 |
| PEO2 | −4.39[b] | −5.66 ± 0.11 | −1.27 | −4.02 ± 0.10 | 0.37 |
| B3MO1 | −4.42[b] | −5.74 ± 0.16 | −1.32 | −4.94 ± 0.08 | −0.52 |
| CPOH | −5.49[b] | −6.87 ± 0.06 | −1.38 | −6.14 ± 0.09 | −0.65 |
| | | Ethers | | | |
| MTHF | −3.34[c] | −5.09 ± 0.13 | −1.74 | −3.99 ± 0.10 | −0.65 |
| DIOX | −5.06[b] | −7.39 ± 0.13 | −2.33 | −5.30 ± 0.16 | −0.24 |
| MPET | −1.69[c] | −2.36 ± 0.11 | −1.69 | −1.60 ± 0.06 | 0.09 |
| EPET | −1.84[c] | −2.88 ± 0.08 | −1.84 | −1.59 ± 0.04 | 0.25 |
| | | overall average | −1.59 | | −0.14 |

[a] Experimental data from ref 68. [b] Experimental data from ref 82. [c] Experimental data from ref 67.

CPOH. In the calculations without any pair-specific LJ parameters, the error in the calculated value of $\Delta G_{\mathrm{hyd}}$ for MTHF is −1.74 kcal/mol, the error in the calculated value for CPOH is −1.38 kcal/mol, and the largest error is −2.33 kcal/mol, obtained for DIOX. Overall, the average error across the whole set of test molecules is −0.14 kcal/mol (rmsd = 0.38 kcal/mol) when pair-specific LJ parameters are included, compared to −1.59 kcal/mol (rmsd = 1.63 kcal/mol) in their absence. Performing a Student's *t* test[71] at a significance level of 0.05 results in rejection of the null hypothesis that the mean error in the $\Delta G_{\mathrm{hyd}}$ values calculated with pair-specific LJ parameters is the same as the mean error in the $\Delta G_{\mathrm{hyd}}$ values without pair-specific LJ parameters ($P$ value $\leq$ 0.0001). From this it can be concluded that the inclusion of pair-specific LJ parameters results in a statistically significant improvement in the reproduction of hydration free energies. It should also be noted that the worst performing of the test set molecules, MTHF and CPOH, both include an atom type for which pair-specific LJ parameters have not been optimized but rather selected by analogy to the corresponding THF atom types. This approach is not necessarily valid, and it is likely that, by optimizing the pair-specific LJ parameters associated with this atom type, some improvement in the calculated value of the MTHF and CPOH hydration free energies could be obtained. It is also worth considering the issue of sampling. As noted above, adequate sampling of conformational space is essential if accurate $\Delta G_{\mathrm{hyd}}$ values are to be obtained for any molecule. It is also something that is increasingly difficult for molecules with increased flexibility, requiring multiple, long simulations. For a molecule such as HEPT, it is extremely unlikely that the entirety of conformational space has been well sampled using the approach outlined above, and the presented values of the hydration free energies should be treated with some caution. For the purpose of this study, however, where the calculations on these longer, more flexible molecules are not targeted at the production of highly accurate hydration free energies, but rather an assessment of whether the pair-specific LJ parameters have resulted in an improvement in the calculated $\Delta G_{\mathrm{hyd}}$ values, these calculations are considered adequate.

When developing optimized force field parameters such as this, it is important to be aware of the risk of overfitting: the situation that occurs when a statistical model describes the data within a training set extremely well, but fails in external test cases. The failure, which occurs when a model possess too many degrees of freedom in relation to the amount of data used for optimization, is often indicative of a model that is not correctly accounting for the underlying physics. In a case such as this study, where 14 pair-specific LJ parameters are fitted to 19 experimental data points, the risk of overfitting is considerable. As a first test for overfitting, the performance of the pair-specific LJ parameters can be compared between the training set and the test set. To do this, a Student's *t* test[71] was performed to assess whether the mean error observed in the training set was significantly different from the mean error observed in the test set; i.e., whether the fitted parameters are having a differential impact on the training versus the test set of molecules, which would indicate overfitting. From this analysis, a $P$ value of 0.3260 was obtained suggesting that the two means may be the same, and it is concluded that there is no significant difference between the mean error observed in the training set and the mean error observed in the test set. Thus, there is no evidence that the pair-specific LJ parameters perform any differently in the training set than they do in the test set. This supports the conclusion that the data is not overfitted. As a second test for overfitting, the modified Akaike Information Criterion ($\mathrm{AIC_C}$)[72] was considered. $\mathrm{AIC_C}$ is a method that can be used to assess the relative information content in competing models of the same data. It works by rewarding accurate reproduction of reference data but penalizing the inclusion of additional parameters. $\mathrm{AIC_C}$ is evaluated via eq 19

$$\mathrm{AIC_C} = 2k + n\ln\left(\frac{\mathrm{RSS}}{n}\right) + \frac{2k(k+1)}{n-k-1} \qquad (19)$$

where $k$ is the number of free parameters, $n$ is the number of observations, and RSS is the residual sum of squares. When comparing models, the model having the lowest $\mathrm{AIC_C}$ score is accepted as the best performing model. Here, there are two competing models: the model without pair-specific LJ parameters, which has no free parameters, and the model with pair-specific LJ parameters, which has 17 free parameters (14 from the original training set, with another 3 added for the test set molecules). Considering all molecules (training set + test set) together, the model without pair-specific LJ parameters has $\mathrm{AIC_C} = 21.70$ and the model with pair-specific LJ parameters has $\mathrm{AIC_C} = -18.40$. This result indicates that the inclusion of pair-specific LJ parameters results in a better model for the calculation of hydration free energies and further supports the conclusion that the model is not overfitted. In theory, it would also be possible to extend this $\mathrm{AIC_C}$ analysis to include the entire body of data used in the development of the CHARMM Drude polarizable force field, not just the solvation free energies. In practice, however, determining the number of free parameters and constructing a RSS with contributions from a variety of different properties would be difficult. What is clear is that the total number of parameters used in each model will be identical, apart from those introduced here, and that both models will give identical results in all areas that do not involve interactions with water. The total $\mathrm{AIC_C}$ values would depend on the magnitude of the contribution to the RSS arising from the additional data points: let us assume that the contribution to the RSS, per data point, would be the same as the average contribution to the RSS, per data point, from the solvation free energy values obtained using the model including pair-specific LJ parameters. If this assumption were correct, then as long as the number of data points increases by more than about 1.3 times the number of parameters, the $\mathrm{AIC_C}$ value for the model including pair-specific parameters will be lower than that of the model without pair-specific LJ parameters.

**3.5. Testing the Need for Pair-Specific LJ Parameters.** The question remains as to whether it is necessary to include pair-specific LJ parameters within the CHARMM Drude polarizable force field for the accurate calculation of hydration free energies. To address this, the pair-specific LJ parameters obtained here were inverted to back-generate a new set of type-specific LJ parameters, as described in the methods section. Using these new LJ parameters, simulations were performed on the bulk neat liquids to calculate thermodynamic properties for a number of alkane and ether molecules. For each of these molecules, the results of these calculations were compared to experimental results, and the results of calculations performed using the standard CHARMM Drude polarizable force field parameters (Table 2). In the initial development of CHARMM Drude polarizable models of small molecules, the reproduction of liquid (or crystal) phase thermodynamic data is considered to be of paramount importance, with parameter optimization performed to yield $V_m$ and $\Delta H_{vap}$ that are both within 2% of the experimental value. As Table 2 shows, this target is almost always achieved. When the corresponding values are calculated using the pair-specific LJ parameters, however, the agreement is considerably worse. Specifically, none of the calculated values are within the 2% target, with the majority of $\Delta H_{vap}$ differing from the experimental target by more than 20%. Overall, using the LJ parameters obtained from the pair-specific LJ parameters, the average error in $V_m$ is 11.2% and the average error in $\Delta H_{vap}$ is −25.0%, compared to average errors of 0.4% and −0.4% in the calculated values of $V_m$ and $\Delta H_{vap}$, respectively, obtained using the standard LJ parameters. Notably, there are systematic differences in the pure solvent properties obtained with the pair-specific parameters, where the $V_m$ values are too large and the $\Delta H_{vap}$ values are all too small. These results, combined with the systematic overestimation of the $\Delta G_{hyd}$ values with the parameters based on the combining rules (Table 1), strongly indicate that the need for additional optimization of the LJ parameters is not associated with limitations in the optimization procedure but rather an inherent limitation in the energy function.

To better quantify the physical underpinnings of the need for the pair-specific LJ parameters, the results of the FEP calculations were analyzed in greater detail. The free energy decomposition approach used to calculate $\Delta G_{hyd}$ (eq 8) allows for the individual contributions to $\Delta G_{hyd}$ due to the WCA-repulsive, WCA-dispersive, and electrostatic interactions to be quantified separately. By examining the change in these contributions upon going from LJ parameters obtained from the combining rules, to pair-specific LJ parameters, a more complete picture can be obtained. The results of this analysis are shown in Table 6 (complete details of the contributions are shown in Table S4 of the Supporting Information). A fascinating trend is revealed: the contribution that is the most affected by the introduction of pair-specific LJ parameters is always associated with the dispersion interaction, with this term always becoming less favorable with the pair-specific LJ parameters. Even with the polar species, the ethers and alcohols, the dispersion term dominates, typically overriding a more favorable electrostatic contribution associated with the pair-specific LJ parameters. These trends allow for several observations. First, the repulsive term, which is dominated by the $1/r^{12}$ portion of the LJ potential, has the smallest contribution. This is reassuring, as this aspect of the LJ treatment of vdW interactions is known to be a fairly poor approximation of a physically more accurate exponential repulsion.[73] While criticism of the $1/r^{12}$ repulsion is still valid, this term does not adversely impact the free energies of aqueous solvation, suggesting that its use in the energy function is not having a significant adverse impact on force field calculations in general. Second, the observation that the electrostatics are not leading to systematic problems validates the inclusion of polarization in the model and suggests that its inclusion is satisfactorily modeling the change in the electronic response of the system in environments of different polarities. Finally, the analysis of the free energy decomposition points to some limitations in the treatment of the dispersive interactions. As the functional form of the dispersive interaction, $\sim 1/r^6$, is physically

***Table 6.*** Variation in the Free Energy Contributions to $\Delta G_{\text{hyd}}$ upon the Introduction of Pair-Specific LJ Parameters (all values in kcal/mol)

| molecule | WCA-repulsion | WCA-dispersion | electrostatic |
|---|---|---|---|
| | | Alkanes | |
| CPEN | −0.16 | 1.18 | −0.01 |
| CHEX | −0.05 | 0.78 | 0.00 |
| ETHA | −0.09 | 0.19 | −0.01 |
| PROP | 0.05 | 0.67 | −0.06 |
| BUTA | −0.14 | 1.01 | −0.05 |
| IBUT | −0.01 | 1.01 | −0.28 |
| NEOP | 0.16 | 1.25 | 0.00 |
| | | Alcohols | |
| MEOH | 0.00 | 0.00 | 0.00 |
| ETOH | −0.13 | 0.58 | −0.17 |
| PRO2 | −0.29 | 0.98 | −0.64 |
| BUO2 | 0.08 | 1.31 | −0.09 |
| PRO1 | −0.20 | 1.03 | −0.62 |
| BUO1 | −0.05 | 1.39 | −0.54 |
| | | Ethers | |
| THF | −0.09 | 1.19 | 0.11 |
| THP | 0.26 | 1.84 | 0.15 |
| DEE | −0.11 | 1.17 | −0.29 |
| DMOE | −0.23 | 1.25 | −0.48 |
| DME | 0.05 | 0.34 | −0.40 |
| MEET | 0.00 | 0.76 | −0.06 |

correct,[73] this indicates that the major limitations arise from the LJ combining rules.

To investigate a possible limitation in the LJ combining rule, the graphical approach of Waldman and Hagler[74] has been applied, focusing on the aliphatic carbon parameters in which the pair-specific parameters only included changes in $\varepsilon$. The plots, which are based on a reduced representation of the change in $\varepsilon_{ij}$ as a function of $\varepsilon_{jj}$ with normalization based on $\varepsilon_{ii}$, the well depth of the water oxygen, are shown in Figure 8. Included are the $\varepsilon_{ij}/\varepsilon_{jj}$ values for the aliphatic carbons based on the data in Table 4 along with curves associated with different types of combining rules. Comparing the pair-specific $\varepsilon$ values obtained in this work to those that would be obtained using either an arithmetic combining rule or the geometric combining rule, eq 3, which is used in CHARMM for the $\varepsilon$ term, shows the limitation in these simple combining rules. The arithmetic mean is clearly inappropriate for $\varepsilon$, as previously discussed,[74] and it is clear that the geometric mean combining rule overestimates the magnitudes of the $\varepsilon$ values required to give an accurate reproduction of experimental data, consistent with the observation of Halgren that "the geometric-mean rule consistently overestimates the well depth for unlike-pair interactions."[75] This leads to the overestimation of the $\Delta G_{\text{hyd}}$ values based on the combining rules (Table 1) and is consistent with the free energy decomposition (Table 6). Applying the combining rules of Waldman and Hagler or of Halgren (Figure 8) results in $\varepsilon$ values that are of smaller magnitude compared to those from the geometric rule, but still too large to reproduce accurately the parameters obtained in this study.

Although none of the tested combining rules are able to reproduce the pair-specific $\varepsilon$ values, the results of the graphical analysis are encouraging. The $\varepsilon$ parameters obtained in this work behave in a very similar manner to those investigated by Waldman and Hagler for the noble gases. They lie on one single curve and, as Waldman and Hagler note, "if there is a valid combination rule *g* that correlates *a*, *b*, and *c*, *then a plot of c/a vs b/a should lie on a single curve.*"[74] This suggests that there should be some combining rule that is able to generate the $\varepsilon$ parameters obtained from the fitting performed in this work. Deriving that combining rule remains a nontrivial task, but an empirical fitting based on the geometric mean rule yields a combining rule (eq 20) that gives an acceptable reproduction of the data shown in Figure 8.

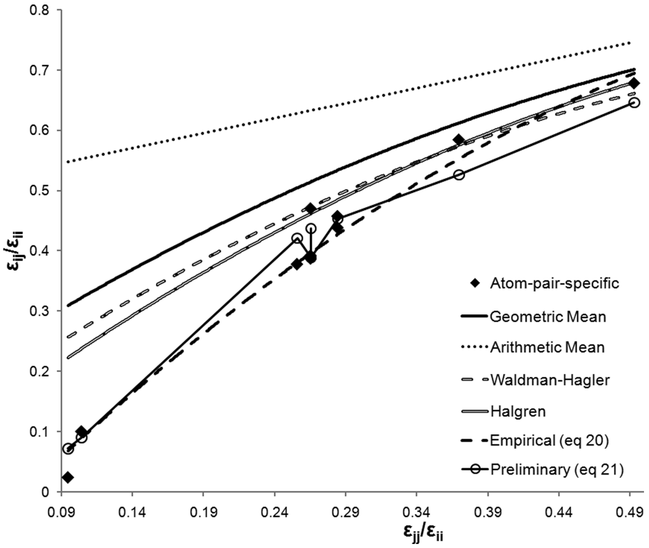


***Figure 8.*** Waldman−Hagler graphical analysis of $\varepsilon_{ij}$ parameter values. Only $\varepsilon_{ij}$ values corresponding to interactions between C atoms and water O atoms are considered. *i* corresponds to the O water atom and *j* to the C atom.

$$\varepsilon_{ij} = 1.6\sqrt{\varepsilon_{ii}\varepsilon_{jj}} - 0.09 \qquad (20)$$

While eq 20 adequately models the data in Figure 8, it has no sound theoretical basis and does not fulfill the basic mathematical requirements of a combining rule.[76] Accordingly, further analysis of the data was performed from which a preliminary combining rule with a more physical basis was empirically determined (eq 21). Based around the $\varepsilon$ combining rule proposed by Halgren,[75] eq 21 also incorporates a term based on the geometric mean rule for $\varepsilon R_{\min}{}^{6}$ as proposed by Waldman and Hagler.[74] The whole expression is then multiplied by an additional term that facilitates an accurate reproduction of the steeper gradient observed for the pair-specific $\varepsilon$ parameters. While this equation is highly preliminary, being specific for only alkane carbons, and unlikely to be the ultimate solution to the problem, it does demonstrate that it is possible to find a combining rule that provides a good representation of the empirically fitted parameters obtained in this work. It also lends further support to the idea that improved combining rules would facilitate an improved force field. Considered in combination with previous studies that have shown that the combining rules used in CHARMM are suboptimal,[77,78] and that the use of alternative combining rules can give improved reproduction of experimental data,[78,79] these results becomes even more persuasive.

$$\varepsilon_{ij} = \left(2 - \frac{2\varepsilon_{ii}\varepsilon_{jj}}{(\varepsilon_{ii} + \varepsilon_{jj})^2}\right)^{0.25}\left[\frac{4\varepsilon_{ii}\varepsilon_{jj}}{(\varepsilon_{ii}^{1/2} + \varepsilon_{jj}^{1/2})^2} - \frac{1}{4}\left(1 - \frac{2R_{\min,ii}^3 R_{\min,jj}^3}{R_{\min,ii}^6 + R_{\min,jj}^6}\right)\right] \quad (21)$$

The inability of available combining rules to treat the present results for the aliphatic carbons is suggested to be associated with the target data used in development of those rules. Combining rules to date have targeted experimental potential energy curves for rare gas homo- and heterodimers. Such data is limited in that it only includes binary interactions of nonpolar atoms whose interactions are dominated by dispersion interactions. The present data are based on complex mixtures of nonpolar and polar molecules, in which significant electrostatic contributions occur. The presence of these contributions is suggested to yield the trend shown in Figure 8; smaller $\varepsilon$ values are required as the value of $\varepsilon$ becomes smaller than that predicted by the standard combining rules. Such small $\varepsilon$ values lead to a decrease in the dispersion contribution to $\Delta G_{\text{hyd}}$, which may be required due to favorable electrostatic contributions on the more polar systems being investigated. While speculative, these results clearly emphasize the importance of the target data in determining an appropriate combining rule for condensed phase studies of polar systems. In the present study this data has been generated on the basis of extremely careful and systematic optimization of LJ parameters initially obtained on the basis of a well-defined set of target data (i.e., based on pure solvent or crystal properties and rare gas interactions) followed by additional optimization to obtain pair-specific LJ parameters to reproduce a second set of well-defined target data (experimental $\Delta G_{\text{hyd}}$ data). The resulting sets of LJ parameters allowed for the development of the preliminary combining rules presented in eqs 20 and 21.

**3.6. Implementing the New Parameters within the CHARMM Drude Polarizable Force Field.** The analysis presented above indicates that the standard combining rule for $\varepsilon$ is not adequate. This problem can be solved by either changing the form of the combining rule or applying the derived pair-specific parameters in the context of the present energy function. Following the former course of action is daunting and would require several steps. First, systematic optimization of the pair-specific LJ parameters would need to be performed in the context of the current combining rules for all the molecules in the force fields for which experimental $\Delta G_{\text{hyd}}$ data are available. Once those values are obtained, a novel combining rule, similar to that in eq 21, would need to be developed, taking into account the full range of molecules in the force field. Once this combining rule is decided upon, new LJ parameters for the entire force field would be required on the basis of the new combining rule, starting with water, through the alkanes and onto the polar molecules and ions. Such a task, while possible, would take several years to complete; to indicate the timeline of such efforts, the first water model for the Drude polarizable force field was published in 2003.[21] The alternative is to apply the pair-specific parameters presented in this study. While this represents a compromise, it is an improvement over the current combining rule based LJ parameters, leading to a better representation of the balance of energetics in bulk systems (e.g., the interior of a protein or lipid bilayer) and in aqueous solution. Such an approach is not unprecedented as Shirts and Pande,[37] for example, have demonstrated (for an additive force field) that it is possible to modify the standard TIP3P water model[80] so as to eliminate the systematic error in hydration free energies without sacrificing the properties of liquid water. In practice, we plan to follow both paths. Over the long-term we anticipate systematically optimizing pair-specific LJ parameters, leading to a new LJ combining rule for $\varepsilon$. In the short term we will extend the small molecule Drude force field to macromolecules using the current combining rule along with the pair-specific LJ parameters. Such an extension to macromolecules is not a trivial process, and it is anticipated that additional limitations in the model will be identified. Corrections to those limitations will then be combined with an improved LJ combination rule to yield a second generation polarizable force field.

## 4. Conclusions

Pair-specific LJ parameters have been developed to describe the interactions between solute heavy atoms and water O atoms. These new parameters yield accurate calculated hydration free energies of alkanes, alcohols, and ethers that provide a good reproduction of experimental reference values. The changes introduced are small in magnitude relative to the LJ parameters obtained using the standard CHARMM parameter combining rules, with the calculated results highly sensitive to these small magnitude changes. They have also been implemented in a hierarchical fashion beginning from the alkanes, and a parametrization protocol has been developed. This will allow for the addition of pair-specific LJ parameters to new functional groups as they are added to CHARMM Drude polarizable force field, in a fashion that is as straightforward and systematic as possible.

The LJ parameters developed in this work have also been used to calculate hydration free energies for a test set of alkane, alcohol, and ether molecules not considered as part of the parametrization process. In these cases, the new parameters yield an acceptable reproduction of experimental properties that is significantly improved compared to that obtained with the combining rule based LJ parameters. This suggests that the pair-specific LJ parameters are broadly transferable across the alkane, alcohol, and ether molecules.

The pair-specific LJ parameters were also used to generate (via the inverse of the standard CHARMM combining rules) a new set of LJ parameters for use in liquid phase calculations of alkane and ether molecules. These parameters were found to give significant, systematic errors in the calculated values of $V_{\text{m}}$ and $\Delta H_{\text{vap}}$. This result suggests that it will not be possible, within the existing framework of the CHARMM Drude polarizable force field, to find a single set of LJ parameters capable of producing both liquid phase thermodynamic data and hydration free energies in good agreement with experimental results.

The systematic optimization of pair-specific LJ parameters in the present study allowed for additional observations to be made. Decomposition of the calculated $\Delta G_{\text{hyd}}$ results exploiting the WCA free energy methodology (eq 8) allowed for the identification that the impact of the pair-specific LJ parameters was on the dispersion term. This result indicates the utility of the treatment of the repulsive aspect of the vdW interactions using the $1/r^{12}$ term and the suitability of the treatment of electronic polarizability using the classical Drude oscillator model. It also indicates limitations in the LJ combining rule leading to the overestimation of the free energies of solvation. This limitation was investigated in the context of the aliphatic carbons and a systematic difference between LJ parameters from the geometric combining rule used in CHARMM (eq 3) as well as other published combining rules for $\varepsilon$. On the basis of this difference, new combining rules were proposed. These rules, while preliminary, indicate that improvements in the treatment of the vdW interactions in empirical force fields are possible, although significant additional work will be required to achieve such a goal.

**Acknowledgment.** The authors acknowledge financial support from the NIH (GM051501 and GM072558) and computational support from the D.O.D. High Performance Computing, the Pittsburgh Supercomputing Center, and the NSF/TeraGrid computational resources.

**Supporting Information Available:** Methods and results for alkane, alcohol, and ether gas phase heterodimer interactions with water molecules; full details of contributions to $\Delta G_{\mathrm{hyd}}$ obtained from WCA decomposition within FEP calculations. This information is available free of charge via the Internet at http://pubs.acs.org

## References

(1) Macleod, N. A.; Butz, P.; Simons, J. P.; Grant, G. H.; Baker, C. M.; Tranter, G. E. *Isr. J. Chem.* **2004**, *44*, 27.

(2) Macleod, N. A.; Butz, P.; Simons, J. P.; Grant, G. H.; Baker, C. M.; Tranter, G. E. *Phys. Chem. Chem. Phys.* **2005**, *7*, 1432.

(3) Freddolino, P. L.; Arkhipov, A. S.; Larson, S. B.; McPherson, A. *Schulten K Structure* **2006**, *14*, 437.

(4) Wlodek, S. T.; Clark, T. W.; Scott, L. R.; McCammon, J. A. *J. Am. Chem. Soc.* **1997**, *119*, 9513.

(5) Snow, C. D.; Nguyen, N.; Pande, V. S.; Grubele, M. *Nature* **2002**, *420*, 102.

(6) Banavali, N. K.; Huang, N.; MacKerell, A. D., Jr. *J. Phys. Chem. B* **2006**, *110*, 10997.

(7) MacKerell, A. D., Jr. *J. Comput. Chem.* **2004**, *25*, 1584.

(8) Baucom, J.; Transue, T.; Fuentes-Cabrera, M.; Krahn, J. M.; Darden, T. A.; Sagui, C. *J. Chem. Phys.* **2004**, *121*, 6998.

(9) Babin, V.; Baucom, J.; Darden, T. A.; Sagui, C. *J. Phys. Chem. B* **2006**, *110*, 11571.

(10) Harder, E.; Kim, B. C.; Friesner, R. A.; Berne, B. J. *J. Chem. Theory Comput.* **2005**, *1*, 169.

(11) Dougherty, D. A. *Science* **1996**, *271*, 163.

(12) Reddy, A. S.; Sastry, G. N. *J. Phys. Chem. A* **2005**, *109*, 8893.

(13) Gallivan, J. P.; Dougherty, D. A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9459.

(14) Wintjens, R.; Liévin, J.; Rooman, M.; Buisine, E. *J. Mol. Biol.* **2000**, *302*, 395.

(15) Tsou, L. K.; Tatko, C. D.; Waters, M. L. *J. Am. Chem. Soc.* **2002**, *124*, 14917.

(16) Zacharias, N.; Dougherty, D. A. *Trends Pharmacol. Sci.* **2002**, *23*, 281.

(17) Aschi, M.; Mazza, F.; Di Nola, A. *J. Mol. Struct. (Theochem)* **2002**, *587*, 177.

(18) Lopes, P. E. M.; Roux, B.; MacKerell, A. D., Jr. *Theor. Chem. Acc.* **2009**, *124*, 11.

(19) Ma, B. Y.; Lii, J. H.; Allinger, N. L. *J. Comput. Chem.* **2000**, *21*, 813.

(20) Maple, J. R.; Cao, Y.; Damm, W.; Halgren, T. A.; Kaminski, G. A.; Zhang, L. Y.; Friesner, R. A. *J. Chem. Theory Comput.* **2005**, *1*, 694.

(21) Lamoureux, G.; MacKerell, A. D., Jr.; Roux, B. *J. Chem. Phys.* **2003**, *119*, 5185.

(22) Patel, S.; Brooks, C. L., III. *J. Comput. Chem.* **2004**, *25*, 1.

(23) Patel, S.; MacKerell, A. D., Jr.; Brooks, C. L., III. *J. Comput. Chem.* **2004**, *25*, 1504.

(24) Drude, P. *The Theory of Optics*; Green: New York, 1902.

(25) Lamoureux, G.; Harder, E.; Vorobyov, I. V.; Roux, B.; MacKerell, A. D., Jr. *Chem. Phys. Lett.* **2006**, *418*, 245.

(26) Vorobyov, I. V.; Anisimov, V. M.; MacKerell, A. D., Jr. *J. Phys. Chem. B* **2005**, *109*, 18988.

(27) Anisimov, V. M.; Vorobyov, I. V.; Roux, B.; MacKerell, A. D., Jr. *J. Chem. Theory Comput.* **2007**, *3*, 1927.

(28) Lopes, P. E. M.; Lamoureux, G.; Roux, B.; MacKerell, A. D., Jr. *J. Phys. Chem. B* **2007**, *111*, 2873.

(29) Vorobyov, I.; Anisimov, V. M.; Greene, S.; Venable, R. M.; Moser, A.; Pastor, R. W.; MacKerell, A. D., Jr. *J. Chem. Theory Comput.* **2007**, *3*, 1120.

(30) Baker, C. M.; MacKerell, A. D., Jr. *J. Mol. Model.* **2010**, *16*, 567.

(31) Lopes, P. E. M.; Lamoureux, G.; MacKerell, A. D., Jr *J. Comput. Chem.* **2009**, *30*, 1821.

(32) Harder, E.; Anisimov, V. M.; Whitfield, T.; MacKerell, A. D., Jr.; Roux, B. *J. Phys. Chem. B* **2008**, *112*, 3509.

(33) Zhu, X.; MacKerell, A. D., Jr. *J. Comput. Chem.* In press.

(34) Anisimov, V. M.; Lamoureux, G.; Vorobyov, I. V.; Huang, N.; Roux, B.; MacKerell, A. D., Jr. *J. Chem. Theory Comput.* **2005**, *1*, 153.

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

(36) Xu, Z.; Luo, H. H.; Tieleman, P. *J. Comput. Chem.* **2006**, *28*, 689.

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

(38) Deng, Y.; Roux, B. *J. Phys. Chem. B* **2004**, *108*, 16567.

(39) Oostenbrink, C.; Villa, A.; Mark, A. E.; van Gunsteren, W. F. *J. Comput. Chem.* **2004**, *25*, 1656.

(40) Mobley, D. L.; Dumont, E.; Chodera, J. D.; Dill, K. A. *J. Phys. Chem.* **2007**, *111*, 2242.

(41) Hunter, C. A.; Sanders, J. K. M. *J. Am. Chem. Soc.* **1990**, *112*, 5525.

(42) Baker, C. M.; Grant, G. H. *J. Chem. Theory Comput.* **2006**, *2*, 947.

(43) Baker, C. M.; Grant, G. H. *J. Chem. Theory Comput.* **2007**, *3*, 530.

(44) Shirts, M. R.; Pitera, J. W.; Swope, W. C.; Pande, V. S. *J. Chem. Phys.* **2003**, *119*, 5740.

(45) Hess, B.; van der Vegt, N. F. A. *J. Phys. Chem. B* **2006**, *110*, 17616.

(46) Wang, J.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049.

(47) Kaminski, G.; Duffy, E. M.; Matsui, T.; Jorgensen, W. L. *J. Phys. Chem.* **1994**, *98*, 13077.

(48) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225.

(49) Geerke, D. P.; van Gunsteren, W. F. *ChemPhysChem* **2006**, *7*, 671.

(50) Ben-Naim, A.; Marcus, Y. *J. Chem. Phys.* **1987**, *81*, 2016.

(51) Wolfenden, R. *Biochem.* **1978**, *17*, 201.

(52) Morgantini, P.-Y.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 6057.

(53) Meng, E. C.; Caldwell, J. W.; Kollman, P. A. *J. Phys. Chem.* **1996**, *100*, 2367.

(54) Ding, Y.; Bernardo, D. N.; Krogh-Jespersen, K.; Levy, R. M. *J. Phys. Chem.* **1995**, *99*, 11575.

(55) Rizzo, R. C.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1999**, *121*, 4827.

(56) Chen, I.; Yin, D.; MacKerell, A. D., Jr. *J. Comput. Chem.* **2002**, *23*, 199.

(57) Davis, J. E.; Warren, G. L.; Patel, S. *J. Phys. Chem. B* **2008**, *112*, 8298.

(58) Rick, S. W.; Berne, B. J. *J. Am. Chem. Soc.* **1996**, *118*, 672.

(59) Kollman, P. *Chem. Rev.* **1993**, *93*, 2395.

(60) Weeks, J. D.; Chandler, D.; Andersen, H. C. *J. Chem. Phys.* **1971**, *54*, 5237.

(61) Lagüe, P.; Pastor, R. W.; Brooks, B. R. *J. Phys. Chem. B* **2004**, *108*, 363.

(62) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*, 1st ed.; Oxford University Press: New York, 1987; pp 64−65.

(63) Brooks, B. R.; Brooks, C. L., III; MacKerell, A. D., Jr.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. *J. Comput. Chem.* **2009**, *30*, 1545.

(64) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327.

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

(66) Wessa, P. Free Statistics Software, version 1.1.23-r5; Office for Research Development and Education. http://www.wessa.net (accessed Feb 2010).

(67) Michieland, L.; Bacilieri, M.; Kaseda, C.; Moro, S. *Bioorg. Med. Chem.* **2008**, *16*, 5733.

(68) Ooi, T.; Oobatake, M.; Némethy, G.; Scherage, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 3086.

(69) Cabani, S.; Gianni, P.; Mollica, V.; Lepori, L. *J. Solution Chem.* **1981**, *10*, 563.

(70) Wolfenden, R.; Andersson, L.; Cullis, P. M.; Southgate, C. C. B. *Biochem.* **1981**, *20*, 849.

(71) Student. *Biometrika* **1908**, *6*, 1.

(72) Akaike, H. *J. Econometrics* **1981**, *16*, 3.

(73) Stone, A. J. *The Theory of Intermolecular Forces*, 1st ed.; Oxford University Press: Oxford, United Kingdom, 1997; pp 157−158.

(74) Waldman, M.; Hagler, A. T. *J. Comput. Chem.* **1993**, *14*, 1077.

(75) Halgren, T. A. *J. Am. Chem. Soc.* **1992**, *114*, 7827.

(76) Khalaf Al-Mata, A.; Rockstraw, D. A. *J. Comput. Chem.* **2003**, *25*, 660.

(77) Delhommelle, J.; Millie, P. *Mol. Phys.* **2001**, *99*, 619.

(78) Song, W.; Rossky, P. J.; Maroncelli, M. *J. Chem. Phys.* **2003**, *119*, 9145.

(79) Ewig, C. S.; Thatcher, T. S.; Hagler, A. T. *J. Phys. Chem. B* **1999**, *103*, 6998.

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

(81) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2005**, *1*, 1133.

(82) Rizzo, R. C.; Aynechi, T.; Case, D. A.; Kuntz, I. D. *J. Chem. Theory Comput.* **2006**, *2*, 128.

CT9005773

Journal of Chemical Theory and Computation

# Assessing the Performance of Popular Quantum Mechanics and Molecular Mechanics Methods and Revealing the Sequence-Dependent Energetic Features Using 100 Tetrapeptide Models

Jinliang Jiang,[†] Yanbo Wu,[†] Zhi-Xiang Wang,*[,†] and Chun Wu*[,‡]

*College of Chemistry and Chemical Engineering, Graduate University of Chinese Academy of Sciences, Beijing, 100049, China, and Department of Chemistry and Biochemistry, University of California—Santa Barbara, Santa Barbara, California 93106*

Received January 6, 2010

**Abstract:** A reasonable description of the conformation energies of each of the amino acids is crucial for modeling protein structures and dynamics. We here used 20 tetrapeptides (ACE-ALA-X-ALA-NME, X = one of 20 amino acids) in 5 conformations (right-handed helix ($\alpha_R$), left-handed helix ($\alpha_L$), $\beta$-sheet ($\beta$), antiparallel $\beta$-sheet ($\beta_a$), and polyproline II (PPII)) as structural models to investigate the relative conformation energies at the MP2/cc-pVTZ//B3LYP/6-31G** level. The results indicate that the energetic pattern (the order and the energy gap) of the five conformations bears certain resemblances among the amino acids in the same class but is quite different among the amino acids in the different classes (e.g., hydrophobic, aromatic, polar and charged classes). The MP2 energies are then used to statistically evaluate the overall performance of various methods including density functional methods (M05-2X, PBE, and B3LYP), semiempirical methods (AM1, PM3, and PM3MM), empirical polarizable force fields (AMOEBA and AMBER), additive force fields (AMBER, CHARMM, GROMOS, OPLS-AA), and united-atom force fields (AMBERUA and GROMOS). In general, M05-2X obviously outperforms PBE and B3LYP. The semiempirical methods are not able to reach the accuracy as expected. Some of the additive force fields are more accurate than the semiempirical methods. The AMOEBA polarizable force field has accuracy comparable with (or better than) the B3LYP and PBE methods. AMBER99, OPLS-AA, CHARMM27 (excluding $\alpha_L$), and AMBERUA (excluding $\alpha_L$) reach reasonable accuracy. However, further improvements, in particular on left-handed helical ($\alpha_L$) and some residues such as Pro, Asp, and Glu, are necessary.

## 1. Introduction

A reliable description of conformation energy is crucial for modeling structures and dynamics of biological systems (e.g., proteins, RNA and DNA). To obtain conformation energy accurate enough for biological applications, the weak nonbonding interactions must be properly taken into account. This requires a high degree of electron correlation energy to be accounted. The quantum mechanics (QM) CCSD(T) approach,[1] coupled-cluster with single and double and perturbative triple excitations, is often considered to be reliable in describing such weak nonbonding interactions. But CCSD(T) is extremely time-consuming at the scale of $O(N^7)$ where $N$ is the number of basis functions, which limits its applications to large molecular systems. The second-order Møller−Plesset perturbation (MP2)[2] method is much less expensive (at a scale of $O(N^5)$) than CCSD(T) and can reach reasonable accuracy in describing the nonbonding interac-

* To whom correspondence should be addressed. E-mail: zxwang@gucas.ac.cn (Z.X.W.); cwu@chem.ucsb.edu (C.W.).

[†] Chinese Academy of Sciences.

[‡] University of California—Santa Barbara.

10.1021/ct100008q © 2010 American Chemical Society
Published on Web 03/01/2010

tions. The methods based on density functional theory (DFT) such as the widely used B3LYP[3,4] and PBE[5] functional have a better scale of $O(N^4)$ than the former two molecular-orbital-based methods, but they are generally not reliable in accounting for nonbonding interactions. Recently, Zhao and Truhlar[6] have developed the M05-2X/M06-2X functionals that account for medium-range correlation energies and thus provide a better description of nonbonding interactions.[7,8] A further compromise between accuracy and computational cost is provided by semiempirical methods (e.g., AM1,[9] PM3, and PM3MM[10,11]). These methods are simplified versions of Hartree−Fock theory by using empirical parameters derived from experimental data, which bring the possibility to study large molecules (up to hundreds of atoms). However, the quality of the DFT and the semiempirical methods in estimating conformation energies is unclear.

Large sizes of biological molecules (e.g., protein, RNA, and DNA) and long time scales of dynamic processes of biological systems (e.g., protein and RNA folding) severely limit the applications of the QM methods. Alternatively, molecular mechanics (MM) modeling provides a tractable approach to describe large biological molecules and make it possible to study the dynamics of biological processes. MM methods describe molecular systems at the atom or united-atom particle level (e.g., aliphatic hydrogen atoms are combined to the connected carbons). Instead of solving the time-consuming Schrödinger equation, MM methods simplify the total potential energy of a molecular system into the sum of several physically meaningful interaction terms (harmonic bond stretching, angle bending, Fourier series for torsion distortion, and Coulomb and Lennard-Jones terms for nonbonding interactions). Anharmonic and cross-terms may be added to improve the accuracy of the force fields.[12] The function of the potential energy and the involved parameters constitute a so-called force field. The force field is the cornerstone of any MM molecular modeling.

Considerable research efforts have been dedicated to developing reliable force fields. The conventional force fields, which have been widely used in studying biological systems, include AMBER,[13] CHARMM,[14] GROMOS,[15,16] and OPLS.[17,18] One of the major defects for the conventional force fields is using fixed partial charges to account for the electrostatic interactions, which neglects the atomic charge changes due to intra- and intermolecular polarization effects. As a consequence, developments of polarizable force fields have been pursued as the next generation of force fields. On the basis of their conventional framework, AMBER, CHARMM, GROMOS, and OPLS have further been developed to implicitly or explicitly include polarization effects.[19–26] In addition to those, the ABEEM developed by Yang's group,[27] AMOEBA developed by the Ponder group,[28] and SIBFA force field developed by Gresh et al.[29] are polarizable force fields for biological systems. In spite of the progress made in developing polarizable force fields, polarizable force fields have not been widely used for studying biological systems due to an elevated computational cost and lack of benchmarking studies to show the benefits.

Force fields were often parametrized to fit the geometric and energetic data of small model molecules from experiments and QM calculations. Although the force fields developed by different groups use very similar energy functions, the parameters may differ significantly due to the different parametrization strategies. For example, the AMBER force fields obtained atomic partial charges by fitting to the QM electrostatic fields of model molecules,[30,31] while OPLS-AA and GROMOS derived the charges by molecular dynamics (MD) simulations to reproduce the experimental data of model molecules.[15,18] Thus, the empirical nature of force fields and the variations between different force fields make it necessary to benchmark them. Although the aspects for a sufficient benchmark remain under debate (e.g., energetics versus thermodynamic properties), as an important aspect, it has been widely adopted to directly compare MM energies/structures of model molecules or larger systems with those obtained by high quality QM calculations.

During the past decades, a large number of QM calculations on the small molecules that may be regarded as model units for proteins have been reported.[32–53] Böehm et al.[54] and Gould et al.[55] independently show that AMBER force fields overestimate the stability of the $C_7$ conformation of alanine and glycine dipeptides when compared with their QM results. Beachy et al.[56] optimized 10 conformers of alanine tetrapeptides (ACE-(Ala)$_3$-NME) at the HF/6-31G** level, and the relative conformation energies at the level of local MP2 (LMP2) with the basis set of cc-pVTZ were used to evaluate the popular force fields AMBER (AMBER3, AMBER4.1, and AMBER94), CHARMM (CHARMM19 and CHARMM22), and OPLS (OPLS-AA(2,2), OPLS/A-UA(2,8), OPLS-UA(2,2)), and GROMOS. Their results showed that OPLS-AA(2,2) is the best force field in terms of structure and relative conformation energies. The 10 alanine tetrapeptides were then used by Gresh et al.[29] to evaluate their SIBFA force field which explicitly takes polarization into account via multipole interactions. They showed that the relative energies calculated at the LMP2/6-311G** level could be reproduced by their SIBFA force field with a root-mean-square deviation (RMS) of about 1.3 kcal/mol. Recently, Kaminsky and Jensen[57] calculated dipeptide conformational energies of four amino acids (Gly, Ala, Ser, and Cys) using different QM methods and MM force fields. They found that the B3LYP/6-31G** calculations could not reproduce all the minima found at the MP2/aug-cc-pVDZ(MP2) level, but for the minima that actually exist on the B3LYP potential energy surface, the geometries and relative energies are in good agreement with the MP2 results. For the polarizable force fields, they found that the AMOEBA polarizable force field performs as well as the B3LYP method for ~80% of the conformations but produces ~20% artificial energy minima which are not present on the MP2 energy surface. The fixed charge force fields were only able to reproduce the geometries of approximately half of the conformations, and OPLS_2005 force fields (slightly modified version of the OPLS[58] force fields in the MacroModel program) perform best among their examined force fields. Some authors also have calculated the infinite long polypeptide chain by DFT methods,[59–64] and the comparison with the force fields[65] showed that all force fields overestimate the stability of the helical conformations except for AM-

BER99/AMBER99SB which satisfactorily reproduce all three helical conformations ($\pi$, α, and $3_{10}$ helix).

These benchmarking studies have contributed greatly to force field developments. Nevertheless, the following limitations motivated the present study: (i) The validations were limited to a few amino acids in the limited secondary structure types. In a force field, three sets of main chain torsion parameters are often used: glycine and proline have their own main chain parameters; the parameters obtained by fitting to the potential energy surfaces of alanine dipeptide or analogues[66,67] are often extended to all remaining amino acids under the assumption of transferability. Yet, the transferability remains elusive, and the errors in these torsion parameters could be one of reasons leading to an imbalance of force fields over the major secondary structure types.[68,69] In addition to the common right-handed helix ($\alpha_R$) and $\beta$ secondary structures, a peptide can also adopt a left-handed helix ($\alpha_L$) and polyproline II conformation (PPII), while the latter conformations are rarely considered in the force field parametrization and assessment. Therefore, evaluations over the complete chemical space of 20 amino acids and over the major secondary structure types could bring us a better understanding of the sequence-dependent conformational energetics and the overall behavior of force fields. (ii) Most of the evaluations were limited to alanine and glycine dipeptides. However, this leads to an additional problem: even if a force field can reproduce the relative energies of dipeptides at different conformations, it does not necessarily imply that the force field is adequate for longer polypeptides because the long-range nonadditive interactions in larger systems that are not present in the dipeptides may play an important role in determining the conformation energy of the longer polypeptides. For example, dipeptides are not able to form intramolecular hydrogen bonds (H-bond) like those in helix secondary structures.

In this study, using tetrapeptides as models, we perform a systemic study to investigate the energetic features of the major conformations of amino acids in the common protein secondary structures at the MP2/cc-pVTZ//B3LYP/6-31G** level and then use the MP2 energies as a "standard" to examine the accuracy of various QM and MM methods (see below for the examined methods). Tetrapeptides are the smallest peptides that can contain H-bonds similar to those in the helix secondary structures. We focus on the five major secondary structures of peptides (i.e the right-handed helix ($\alpha_R$), left-handed helix ($\alpha_L$), $\beta$-sheet ($\beta$), anti-$\beta$ sheet ($\beta_a$), and polyproline II conformation (PPII)). In total, 100 tetrapeptide structures (20 amino acids × 5 conformations) were used in this study. To our knowledge, this is the first time a study of the energetic features of all amino acids in the major secondary structures has been done at levels ranging from MP2/cc-pVTZ to the MM-based molecular mechanics model. These results could provide invaluable information to both method developers and users for future development and method selection.

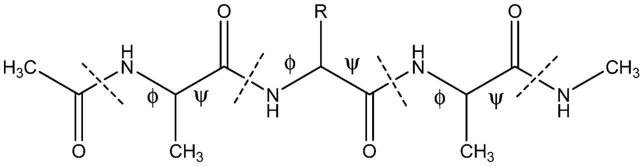


***Figure 1.*** Tetrapeptide model (R = side chain of 20 amino acids).

## 2. Computational Methods

For a tetrapeptide (Figure 1), ACE-ALA-X-ALA-NME (X is one of the 20 amino acids and ACE and NME are respectively acetyl and methylamide groups which cap the tetrapeptide; in comparison with dipeptide models, the tetrapeptide models using two additional ALAs can reduce the errors due to terminal groups in force field assessment), the five typical conformations have the backbone ($\phi/\Psi$) and side chain dihedral angles defined as below. The backbone dihedral angles of the five conformations are the right-handed helix ($\alpha_R$; $\phi = -57.0°$, $\Psi = -47°$), left-handed helix ($\alpha_L$; $\phi = 57.0°$, $\Psi = 47°$), $\beta$-sheet ($\beta$; $\phi = -119.0°$, $\Psi = 113.0°$), anti-$\beta$ sheet ($\beta_a$; $\phi = -140.0°$, $\Psi = 135°$), and polyproline II conformation (PPII; $\phi = -79.0°$, $\Psi = 150.0°$). These ($\phi/\Psi$) angles are applied to the three sets of backbone $\phi/\Psi$ pairs of the tetrapeptides. The rotamer library developed by Dunbrack's group,[70] was used to determine the side chain dihedral angles (except for $\chi_3$ and $\chi_4$ of proline, which are not available from the rotamer library and were obtained from the geometry optimization at B3LYP/6-31G** with $\chi_1$ and $\chi_2$ fixed to the library values). Given the main-chain dihedral angles, the side chain dihedral angles were chosen to be the values in the most populated rotamers. The dihedral angles for the five conformations of each amino acid are provided in the Supporting Information (Table S1 in Supporting Information A (SIA)). All the QM and MM geometric optimizations in the gas phase were carried out with the backbone and side chain dihedral angles fixed to the predefined values. The reason for using these restraints is to prevent the geometric optimizations from producing structures that rarely exist in the peptide/protein structures in aqueous solution (for example, the $C_{7eq}$ conformation is the most stable conformation of alanine dipeptide, but it is rarely seen in protein structures) or from producing divergent structures under different methods, which make the comparisons of conformational energies inconsistent. In other words, the restraints of dihedral angles make it possible to focus our benchmarking on the common protein secondary structures for different methods. The solvation effect is critical in determining protein structures;[71] its influence on the benchmark is under investigation and will be reported in the future.

The structures of the tetrapeptides were optimized at either B3LYP/6-31G** (the optimized structures are drawn in Figure S1 of SIA) or M05-2X/6-31G** levels. The single point energies were then obtained at MP2/cc-pVTZ. Because the two sets of MP2 energies are very close, which is due to the restraints used in the geometry optimizations leading to very similar structures, we only present the data set with the B3LYP/6-31G** optimized structures in the main text, and the other set of results are provided in Supporting Information B (SIB).

The MP2/cc-pVTZ energies are used as the "standard" values to evaluate the performance of all other methods. The examined QM methods include M05-2X/cc-pVTZ//M05-2X/6-31G**, M05-2X/6-31G**//M05-2X/6-31G**, PBE/cc-pVTZ//PBE/6-31G**, PBE/6-31G**//PBE/6-31G**, B3LYP/cc-pVTZ//B3LYP/6-31G**, B3LYP/6-31G**//B3LYP/6-31G**, AM1//AM1, PM3//PM3, and PM3MM//PM3MM (PM3 + the optional molecular mechanics correction for HCON linkages), where the calculation levels behind "//" indicate the levels used in the structural optimizations. All QM calculations were carried out by using the Gaussian 03 program.[72] Note that the M06-2X functional is not available in Gaussian 03 and should be studied in future work. MM calculations were carried out with the Tinker program[73] using different force fields, including AMOEBA,[28] AMBER94,[13] AMBER96,[74] AMBER99,[30] CHARMM27,[14] OPLS-AA,[17] and OPLS-AA/L.[18] AMBER03,[31] AMBER99SB,[75] AMBEREP,[19] AMBERPOL,[20] and AMBERUA[76] were carried out using the Amber 9 package.[77] GROMOS96 force fields (for versions G43b1, G45a3,[15] and G53a6,[16] COOH was used as C-terminal due to lack of NME) were calculated using the GROMACS 3.3 simulation package.[78] The MM energies were obtained on the basis of the reoptimized structures at the corresponding level. A dielectric constant of 1.0 and an infinite cutoff for Lennard-Jones interactions was used in MM calculations. All conformation energies relative to the $\alpha_R$ conformation are provided in Tables S2 of SIA and the Table S1 of SIB).

To statistically evaluate the performance of the examined methods, two types of root-mean-square deviations (RMS) of conformation energies were calculated either for each amino acid, averaged over the five conformations (RMS), or for each conformation type, averaged over 20 amino acids (RMS-C). The first type of RMS is calculated by using eqs 1−3:

$$\mathrm{RMS} = \sqrt{\frac{\sum_{i=1}^{n} (\mathrm{error})^2}{n}} \tag{1}$$

$$\mathrm{error} = E_{ai} - E_{bi} + E_c \tag{2}$$

$$E_c = \frac{\sum_{i=1}^{n} (E_{bi} - E_{ai})}{n} \tag{3}$$

where $n$ is the total number of the conformations (i.e., $n = 5$), $E_{bi}$ and $E_{ai}$ are respectively the relative energies (i.e., setting the energy of $\alpha_R$ to be zero) of the reference method (i.e., MP2) and a given method, error is a signed error using the MP2 energy as the "true" value, and $E_c$ is a constant to minimize the rms for each amino acid type, which fixes the issue that, if the relative energies are defined relative to a given conformation, the rms values will depend on the reference conformation. As indicated by eqs 2 and 3, the use of $E_c$ is actually equivalent to using the mean value of all conformations as the reference. The second type of rms is the signed rms-C of each conformation type averaged over 20 amino acids for a given method and calculated by using eqs 4 and 5:

$$\mathrm{RMS\text{-}C} = \sqrt{\frac{\sum_{j=1}^{m} (E_{aj} - E_{bj} + E_c)^2}{m}} \tag{4}$$

$$\mathrm{SIGN} = \mathrm{sign}\left(\frac{\sum_{j=1}^{m} E_{aj} - E_{bj} + E_c}{m}\right) \tag{5}$$

where $m$ is the total number of the amino acid type (i.e., $m = 20$), $E_{bj}$ and $E_{aj}$ are respectively the relative energies (i.e., setting the energy of $\alpha_R$ to be zero) of the MP2 and a given method for a particular conformation, and $E_c$ is the energy offset obtained from eq 2 in minimizing the RMS for each amino acid type. The reason for not directly using the $E_c$ obtained from eq 4 in minimizing RMS-C for each conformation type (using 20 amino acids) is that such an energy reference $E_c$ should not depend on conformation type. SIGN is determined from eq 5, which determine the sign of the averaged signed error over the 20 amino acids. On the basis of the definition of RMS/RMS-C, one can see that the RMS/RMS-C can provide statistical information of the performance of a given method on an individual amino acid over all five conformations/on individual conformations over all 20 amino acids, respectively.

In addition, RMS and RMS-C are also calculated on the basis of four conformations (i.e., excluding the left-handed helix conformation), because the $\alpha_L$ conformation is only adopted by short peptides and is rarely presented in protein structure modeling. To distinguish them from those calculated over all five conformations, we refer to them as RMS-N$\alpha_L$ and RMS-C-N$\alpha_L$, where N$\alpha_L$ is the abbreviation for "not including $\alpha_L$".

To evaluate the overall performance of the examined methods, the means ($\mu$) of RMS/RMS-N$\alpha_L$ were calculated as the averages over 20 amino acids by taking five/four conformations into account. The means ($\mu$) of unsigned RMS-C/RMS-C-N$\alpha_L$ were calculated as the averages over five/four conformations by taking 20 amino acids into account. The standard deviations ($\sigma$) of RMS/RMS-N$\alpha_L$ and unsigned RMS-C/RMS-C-N$\alpha_L$ were calculated correspondingly, which provide information on whether a given method has a balanced performance on 20 amino acids or on five/four conformations.

## 3. Results and Discussion

For brevity, we use the amino acid name to refer the whole tetrapeptide hereafter. The energies at the MP2/cc-pVTZ//B3LYP/6-31G** level relative to the α-helix conformation are listed in Table 1 and are plotted in Figure 2.

From left to right in Figure 2, we order the results by following the common classifications of 20 amino acids: hydrophobic (Pro-Met), aromatic (Phe-Trp), polar (Cys-Gln), and charged (Asp-Arg) classes. As expected, the pattern in terms of energy order and gaps of the five conformations

***Table 1.*** Relative Energies (kcal/mol; with reference to the $\alpha_R$ conformation) of the 100 Tetrapeptide Structures at the MP2/cc-pVTZ//B3LYP/6-31G** Level

| | $\alpha_R$ | $\alpha_L$ | $\beta$ | $\beta_a$ | PPII |
|---|---|---|---|---|---|
| Pro | 0.00 | 2.24 | 4.26 | 5.29 | 1.76 |
| Gly | 0.00 | 1.55 | 3.67 | 1.74 | 2.85 |
| Ala | 0.00 | 2.23 | 3.44 | 1.01 | 2.91 |
| Val | 0.00 | 3.61 | 2.15 | 0.18 | 3.17 |
| Leu | 0.00 | 3.24 | 3.37 | 0.79 | 3.01 |
| Ile | 0.00 | 4.15 | 1.97 | 0.13 | 2.27 |
| Met | 0.00 | −0.12 | 3.05 | 0.87 | 1.53 |
| Phe | 0.00 | −0.62 | 0.55 | −0.91 | 2.69 |
| Tyr | 0.00 | −0.43 | 0.37 | −0.97 | 2.51 |
| His | 0.00 | 0.05 | 1.28 | 0.13 | 2.32 |
| Trp | 0.00 | −0.19 | 2.54 | −0.27 | 2.40 |
| Cys | 0.00 | 0.95 | 1.78 | 1.15 | 4.06 |
| Ser | 0.00 | 1.33 | 3.90 | 1.13 | 3.62 |
| Thr | 0.00 | 1.99 | 5.44 | 4.22 | 4.00 |
| Asn | 0.00 | −0.84 | 1.36 | 1.53 | 3.52 |
| Gln | 0.00 | 2.06 | 2.63 | 0.43 | −1.74 |
| Asp | 0.00 | 16.70 | −0.44 | 7.01 | 12.97 |
| Glu | 0.00 | 5.34 | −0.63 | 2.91 | 7.20 |
| Lys | 0.00 | 2.92 | 3.51 | 0.42 | 0.78 |
| Arg | 0.00 | −4.06 | −1.33 | −18.48 | −4.90 |

for each amino acid shows certain similarity within the class (e.g., Gly vs Ala, Phe vs Tyr) and more obvious differences between the classes (e.g., aromatic and charged classes vs hydrophobic and polar classes). The main features are summarized as below: (1) $\alpha_R$ conformations are the lowest in the hydrophobic (except for Met) and polar classes (except for Asn and Gln), but this is not the case in the charged (except for Lys) and aromatic classes. This data may indicate the following: when the interaction between the side chain and backbone is weak, as in the hydrophobic and polar classes, the backbone interactions (e.g., H-bond) dictate the conformation energy, but when the side chain−backbone interaction is strong, as in the charged and aromatic classes, it may overtake the backbone interactions and change the energetic pattern. (2) In the aromatic class, the energy of $\alpha_L$ is close to or less than that of $\alpha_R$ (−0.62, −0.43, −0.19, and 0.05 kcal/mol energy difference for Phe, Tyr, Trp, and His, respectively). This might be caused by the ring−backbone interactions. (3) In the charged class, the energy gaps between different conformations become large except for Lys. For example, the energy gaps between the lowest energy conformation and the highest energy conformation is 7.83, 17.14, and 18.48 kcal/mol for Glu, Asp, and Arg, respectively. In addition, Arg has an extremely favorable $\beta_a$ conformation, at least 13.58 kcal/mol lower than other conformations; this is caused by H-bond interaction between the N−H of the side chain and the C=O of the backbone for $\beta_a$ (Figure S1 in SIA). (4) Although $\alpha_L$ appears to be "mirror" image of $\alpha_R$ in a reduced ribbon representation, their energies are not close to each other. In fact, $\alpha_L$ has a higher energy than $\alpha_R$ for 13 amino acids (except for the whole aromatic class, Met, Asn, and Arg) by at least 1.0 kcal/mol; this can be attributed to the stronger steric effect between C=O groups and side chains in $\alpha_L$ than that between N−H groups and side chains in $\alpha_R$. However, in the case of Arg, $\alpha_L$ has a lower energy than $\alpha_R$ by −4.06 kcal/mol, which might stem from the different side chain torsions in the two conformations. In short, no two amino acids have the exact same conformation energy profile, highlighting the sequence dependent features. From the perspective of secondary structure propensity, the conformation energy profile determines the intrinsic preference toward certain secondary structure types (extended conformation versus helical conformation) for each amino acid. Thus, it is critical for lower level methods to reproduce these energetic signatures of each amino acid as accurately as possible. The errors may lead to the wrong secondary structure propensity for each amino acid.

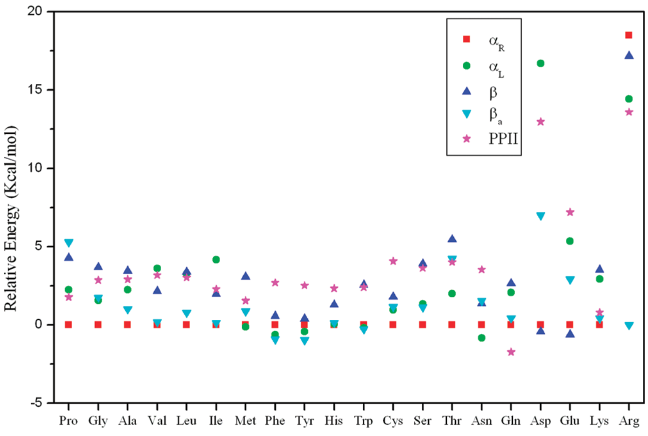


***Figure 2.*** Relative energies of five conformations of 20 amino acids using MP2/cc-pVTZ//B3LYP/6-31G**.

In the following, we use the MP2 energies as the "standard" to assess the other methods. It should be pointed out that the MP2 energies may have a deviation of about 0.5 kcal/mol with respect to the "true" values calculated at the more sophisticated QM level. For brevity, we mainly discuss the overall performance of the examined methods and place the details in the Supporting Information. Tables 2 and 3 list only the signed RMS-C values and the RMS-N$\alpha_L$ values, respectively. The means ($\mu$) and standard deviations ($\sigma$) of the unsigned RMS-C's and the RMS-N$\alpha_L$'s are plotted in Figures 3 and 4 for visualization, respectively. The RMS-C-N$\alpha_L$ values are given in Table S4c of SIA and included in Figure 3. The RMS values are given in Table S3d of SIA but not included in Figure 4 for brevity.

The RMS-C values in Table 2 and its mean in Figure 3 demonstrate the M05-2X functional obviously outperforms the PBE and B3LYP functional in predicting the relative conformation energies. The mean RMS-C values ($\mu$) of M05-2X/cc-pVTZ (0.79 kcal/mol) and M05-2X/6-31G** (0.84 kcal/mol) are substantially less than those of PBE/cc-pVTZ (2.60 kcal/mol), PBE/6-31G** (1.45 kcal/mol), B3LYP/cc-pVTZ (3.24 kcal/mol), and B3LYP/6-31G** (1.90 kcal/mol). This can be attributed to the better description of nonbonding interactions by the M05-2X functional. As shown in Table 2, at both levels of PBE (PBE/6-31G** and PBE/cc-pVTZ), B3LYP (B3LYP/6-31G** and B3LYP/cc-pVTZ), and M05-2X/cc-pVTZ, the $\alpha_R$ and $\alpha_L$ conformations which contain intramolecular H-bonds have positive RMS-C values, while the RMS-C's of extended conformations ($\beta$, $\beta_a$, and PPII) which do not have intramolecular H-bonds are negative. In other words, these DFT methods overestimate the energies of the compact helical conformations but underestimate the

***Table 2.*** Signed RMS-C (kcal/mol) of Each Conformation over 20 Amino Acids for All Considered Methods

| | $\alpha_R$ | $\alpha_L$ | $\beta$ | $\beta_a$ | PPII | mean ($\mu$) | $^1$SD ($\sigma$) |
|---|---|---|---|---|---|---|---|
| M05-2X/cc-pVTZ | 1.22 | 0.62 | −0.52 | −0.79 | −0.78 | 0.79 | 0.24 |
| M05-2X-D$^a$/cc-pVTZ | −1.09 | 0.86 | 0.92 | 0.89 | 0.54 | 0.86 | 0.18 |
| M05-2X/6-31G** | 0.48 | −1.79 | 0.93 | 0.31 | 0.70 | 0.84 | 0.52 |
| M05-2X-D$^a$/6-31G** | −1.81 | −1.97 | 1.54 | 1.24 | 1.59 | 1.63 | 0.25 |
| PBE/cc-pVTZ | 3.22 | 3.18 | −3.02 | −2.66 | −0.91 | 2.60 | 0.87 |
| PBE-D$^a$/cc-pVTZ | 1.21 | 3.16 | −2.55 | −2.05 | 0.51 | 1.90 | 0.94 |
| PBE/6-31G** | 2.19 | 0.95 | −1.85 | −1.69 | 0.58 | 1.45 | 0.60 |
| PBE-D$^a$/6-31G** | 0.48 | 1.08 | −1.39 | −1.19 | 1.31 | 1.09 | 0.32 |
| B3LYP/cc-pVTZ | 3.89 | 4.12 | −3.30 | −3.06 | −1.83 | 3.24 | 0.80 |
| B3LYP-D$^a$/cc-pVTZ | 1.85 | 4.09 | −2.83 | −2.42 | −1.04 | 2.45 | 1.02 |
| B3LYP/6-31G** | 2.75 | 1.87 | −2.11 | −2.08 | −0.68 | 1.90 | 0.68 |
| B3LYP-D$^a$/6-31G** | 0.81 | 1.91 | −1.63 | −1.50 | 0.56 | 1.28 | 0.51 |
| AM1 | 2.11 | 4.68 | −3.20 | −1.74 | −2.83 | 2.91 | 1.02 |
| AM1-D$^a$ | −1.17 | 4.70 | −2.81 | 1.74 | −1.98 | 2.48 | 1.23 |
| PM3 | 5.54 | 8.01 | −3.93 | −3.01 | −7.27 | 5.55 | 1.90 |
| PM3-D$^a$ | 3.48 | 7.99 | −3.54 | −2.44 | −6.39 | 4.77 | 2.08 |
| PM3MM | 5.73 | 8.74 | −4.98 | −3.45 | −6.71 | 5.92 | 1.77 |
| PM3MM-D$^a$ | 3.67 | 8.69 | −4.56 | −2.76 | −5.83 | 5.10 | 2.06 |
| AMOEBA | 1.21 | −1.61 | 1.67 | 1.78 | −1.54 | 1.56 | 0.19 |
| AMBEREP | −4.11 | 10.24 | −1.89 | −2.59 | −2.68 | 4.30 | 3.06 |
| AMBERPOL | −3.17 | 5.91 | 1.29 | −1.45 | −2.54 | 2.87 | 1.67 |
| AMBER94 | −4.60 | 4.80 | 1.45 | 1.81 | −2.22 | 2.98 | 1.43 |
| AMBER96 | 1.04 | 9.55 | −3.15 | −4.20 | −3.09 | 4.21 | 2.86 |
| AMBER99 | −2.86 | 3.78 | 1.90 | −2.28 | −1.02 | 2.37 | 0.93 |
| AMBER99SB | 2.72 | 2.98 | −1.25 | −2.16 | −3.53 | 2.53 | 0.78 |
| AMBER03 | −1.89 | 9.11 | −1.83 | −3.28 | −4.79 | 4.18 | 2.69 |
| CHARMM27 | −2.99 | 14.55 | −4.10 | −2.87 | −6.02 | 6.11 | 4.37 |
| OPLS-AA | 2.60 | 4.58 | −3.20 | −2.57 | −2.96 | 3.18 | 0.74 |
| OPLS-AA/L | 2.54 | 5.70 | −2.31 | −3.02 | −3.97 | 3.51 | 1.23 |
| AMBERUA | −3.50 | 13.86 | −2.54 | −3.32 | −5.54 | 5.75 | 4.17 |
| GROMOS(G43b1) | 4.33 | 10.65 | −5.63 | −3.79 | −4.53 | 5.79 | 2.50 |
| GROMOS(G45a3) | 3.59 | 9.60 | −5.68 | −3.66 | −4.05 | 5.32 | 2.27 |
| GROMOS(G53a6) | 5.14 | 10.36 | −6.51 | −4.06 | −4.87 | 6.19 | 2.23 |

[a] AMBER 99 dispersion energies are applied (see text for details).

energies of the extended conformations with respect to the corresponding MP2 energies. This can be attributed to the fact that the DFT methods (in particular the B3LYP functional) are not able to account for the nonbonding interactions properly (e.g., underestimation of the dispersion and H-bonding interactions in the compact conformations). Similarly, Table 2 can be used to examine the performance of other methods on the individual conformations.

Because the MP2 energies were computed using the cc-pVTZ basis set, one may assume that the cc-pVTZ basis set could give better agreement than the 6-31G** basis set. However, the mean RMS-C values indicate that the cc-pVTZ basis set only marginally improves the agreement of the M05-2X functional from 0.84 kcal/mol (6-31G** basis set) to 0.79 kcal/mol, but it even worsens the agreement of the B3LYP (and PBE) functional from 1.90 (and 1.45) kcal/mol (6-31G** basis sets) to 3.24 (and 2.60) kcal/mol. This can be attributed to the larger basis set superposition error (BSSE) of the 6-31G** basis set than that of the cc-pVTZ basis set. With respect to the cc-pVTZ basis set, the 6-31G** basis set leads to larger BSSE values for more compact conformations than for the extended conformations, which compensates more for the dispersion that is intrinsically underestimated by DFT methods in the compact conformations than in the extended conformations. Because of the defect of the PBE and B3LYP functionals in accounting for the nonbonding interactions, we simply added the MM dispersion energies obtained from AMBER99 calculations to the PBE and B3LYP energies (denoted by adding suffix "D" in the tables and figures); the agreements of the PBE and B3LYP functionals are improved by about 0.4−0.7 kcal/mol; the mean RMS-C's of the cc-pVTZ basis set are reduced from 2.60 and 3.24 kcal/mol to 1.90 and 2.45 kcal/mol for PBE and B3LYP, respectively; the mean RMS-C's of the 6-31G** basis set are decreased from 1.45 and 1.90 to 1.09 and 1.28 kcal/mol for PBE and B3LYP, respectively. This implies that the PBE and B3LYP functional can be moderately improved by adding the Lennard-Jones potential, indicating that the accuracy of DFT methods can be further improved by treating dispersion interactions in a more systematic way, as exemplified by the M05-2X and M06-2X density functionals. Due to the double counting of the medium-range dispersion effect, such corrections worsen the M05-2X performance by about 0.1−0.8 kcal/mol. The standard RMS-C deviations ($\sigma$) pronounce that the M05-2X/cc-pVTZ method ($\sigma = 0.24$ kcal/mol) has more consistent descriptions of the five conformations than do PBE/cc-pVTZ ($\sigma = 0.87$ kcal/mol) and B3LYP/cc-pVTZ ($\sigma = 0.80$ kcal/mol), which is in agreement with the above discussion.

The examined semiempirical methods are less accurate than both DFT methods. Their mean RMS-C's are larger than those of DFT methods (see Table 2). AM1 ($\mu = 2.91$ kcal/mol) outperforms PM3 ($\mu = 5.55$ kcal/mol). The signed RMS-C values of AM1 and PM3 indicate that the semiempirical methods may share the same reasons for their poor performance as the B3LYP methods but with larger errors. The poor performance of PM3 cannot be rescued by adding a MM correction of the pyramidalization of the amide

***Table 3.*** RMS-N$\alpha_L$ (kcal/mol) over the Four Conformations ($\alpha_R$, $\beta$, $\beta_a$, and PPII) of Each Amino Acid for Each Method

| | Pro | Gly | Ala | Val | Leu | Ile | Met | Phe | Tyr | His | Trp | Cys | Ser | Thr | Asn | Gln | Asp | Glu | Lys | Arg | Mean | SD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| M05-2X/cc-pVTZ | 0.83 | 0.98 | 0.84 | 0.83 | 0.77 | 0.77 | 0.68 | 0.99 | 1.05 | 0.88 | 1.15 | 1.05 | 0.77 | 0.67 | 0.74 | 0.77 | 0.30 | 0.85 | 0.84 | 0.92 | 0.83 | 0.17 |
| M05-2X-D[a]/cc-pVTZ | 2.34 | 0.58 | 0.62 | 0.50 | 0.48 | 0.49 | 0.52 | 0.59 | 0.87 | 0.63 | 0.82 | 1.27 | 0.64 | 0.55 | 0.35 | 0.68 | 0.91 | 1.08 | 0.42 | 0.54 | 0.74 | 0.43 |
| M05-2X/6-31G** | 0.34 | 0.50 | 0.14 | 0.38 | 0.44 | 0.33 | 0.25 | 0.30 | 0.58 | 0.22 | 0.26 | 1.11 | 0.34 | 0.46 | 0.27 | 0.48 | 0.79 | 0.40 | 0.41 | 0.40 | 0.42 | 0.21 |
| M05-2X-D[a]/6-31G** | 2.72 | 1.46 | 1.43 | 1.31 | 1.42 | 1.26 | 1.37 | 1.14 | 1.35 | 1.21 | 1.58 | 1.99 | 1.43 | 1.49 | 1.14 | 1.50 | 1.43 | 1.38 | 1.11 | 0.88 | 1.43 | 0.37 |
| PBE/cc-pVTZ | 2.51 | 2.80 | 2.74 | 2.56 | 2.63 | 2.50 | 2.53 | 2.40 | 2.31 | 2.56 | 2.84 | 2.44 | 2.60 | 2.68 | 2.37 | 2.59 | 1.35 | 2.32 | 2.70 | 2.08 | 2.48 | 0.31 |
| PBE-D[a]/cc-pVTZ | 1.12 | 1.62 | 1.54 | 1.59 | 1.61 | 1.45 | 1.52 | 1.86 | 1.82 | 2.12 | 1.68 | 1.31 | 1.36 | 1.59 | 1.50 | 1.34 | 0.28 | 1.03 | 2.15 | 1.94 | 1.52 | 0.40 |
| PBE/6-31G** | 1.86 | 1.94 | 1.84 | 1.60 | 1.66 | 1.56 | 1.65 | 1.51 | 1.47 | 1.62 | 1.84 | 1.59 | 1.70 | 1.57 | 1.55 | 1.89 | 1.18 | 1.91 | 1.70 | 1.64 | 1.66 | 0.18 |
| PBE-D[a]/6-31G** | 1.33 | 1.18 | 0.99 | 0.90 | 0.99 | 0.61 | 0.98 | 1.30 | 1.38 | 1.50 | 1.12 | 0.88 | 0.72 | 0.68 | 0.96 | 1.15 | 0.75 | 0.98 | 1.43 | 1.76 | 1.08 | 0.29 |
| B3LYP/cc-pVTZ | 2.92 | 3.16 | 3.20 | 3.00 | 3.11 | 2.94 | 2.94 | 2.88 | 2.81 | 3.02 | 3.41 | 2.92 | 3.15 | 3.18 | 2.84 | 2.96 | 1.54 | 2.61 | 3.19 | 2.46 | 2.91 | 0.38 |
| B3LYP-D[a]/cc-pVTZ | 1.47 | 1.82 | 1.87 | 1.90 | 1.97 | 1.86 | 1.85 | 2.13 | 2.08 | 2.37 | 2.08 | 1.64 | 1.85 | 2.02 | 1.86 | 1.61 | 0.53 | 1.32 | 2.48 | 2.09 | 1.84 | 0.40 |
| B3LYP/6-31G** | 2.06 | 2.15 | 2.25 | 2.01 | 2.05 | 1.95 | 2.02 | 1.89 | 1.81 | 1.98 | 2.28 | 1.98 | 2.21 | 2.02 | 1.96 | 2.16 | 1.20 | 1.86 | 2.09 | 1.72 | 1.98 | 0.23 |
| B3LYP-D[a]/6-31G** | 1.14 | 1.00 | 1.06 | 1.01 | 1.06 | 0.87 | 1.09 | 1.32 | 1.27 | 1.52 | 1.18 | 0.84 | 0.98 | 0.90 | 1.09 | 1.08 | 0.56 | 0.66 | 1.52 | 1.53 | 1.08 | 0.25 |
| AM1 | 2.20 | 2.39 | 2.14 | 1.91 | 2.31 | 1.95 | 1.50 | 2.62 | 2.56 | 2.52 | 2.80 | 1.62 | 2.17 | 1.62 | 1.37 | 1.67 | 1.61 | 2.59 | 2.14 | 3.90 | 2.18 | 0.57 |
| AM1-D[a] | 1.70 | 1.22 | 1.07 | 1.14 | 1.24 | 1.49 | 1.02 | 1.70 | 1.63 | 1.81 | 1.16 | 0.57 | 1.95 | 0.94 | 0.44 | 1.16 | 1.39 | 1.58 | 1.76 | 4.18 | 1.46 | 0.74 |
| PM3 | 4.85 | 4.63 | 4.84 | 4.85 | 5.48 | 4.79 | 3.90 | 5.24 | 5.22 | 5.19 | 5.45 | 4.42 | 3.82 | 5.38 | 4.17 | 3.56 | 4.88 | 5.13 | 4.83 | 4.69 | 4.77 | 0.54 |
| PM3-D[a] | 2.85 | 3.37 | 3.58 | 3.64 | 4.35 | 3.99 | 2.85 | 4.10 | 4.06 | 4.20 | 3.96 | 3.26 | 2.91 | 4.27 | 3.23 | 2.37 | 4.24 | 3.93 | 4.04 | 4.35 | 3.68 | 0.58 |
| PM3MM | 4.92 | 4.92 | 4.84 | 4.76 | 5.40 | 4.70 | 3.96 | 5.22 | 5.15 | 5.11 | 5.40 | 4.55 | 4.64 | 5.78 | 4.37 | 3.85 | 4.82 | 4.64 | 5.28 | 4.96 | 4.86 | 0.46 |
| PM3MM-D[a] | 2.77 | 3.55 | 3.50 | 3.54 | 4.23 | 3.83 | 2.84 | 4.14 | 4.04 | 4.16 | 3.84 | 3.25 | 3.68 | 4.59 | 3.33 | 2.54 | 4.22 | 3.43 | 4.48 | 4.58 | 3.73 | 0.58 |
| AMOEBA | 3.92 | 0.70 | 0.70 | 1.62 | 0.83 | 1.07 | 0.53 | 0.46 | 0.27 | 1.56 | 1.01 | 0.90 | 0.51 | 1.94 | 1.58 | 0.23 | 2.50 | 2.10 | 1.38 | 0.91 | 1.24 | 0.87 |
| AMBEREP | 2.91 | 1.35 | 0.72 | 0.42 | 0.93 | 0.54 | 1.30 | 0.73 | 0.78 | 0.50 | 1.42 | 1.96 | 0.95 | 1.07 | 1.66 | 0.68 | 1.66 | 2.94 | 0.31 | 1.73 | 1.23 | 0.73 |
| AMBERPOL | 3.92 | 0.58 | 1.29 | 0.99 | 1.23 | 1.40 | 1.38 | 1.12 | 1.00 | 0.70 | 1.91 | 1.89 | 1.00 | 1.56 | 1.21 | 1.40 | 1.44 | 2.96 | 0.76 | 2.40 | 1.51 | 0.78 |
| AMBER94 | 4.03 | 1.37 | 2.16 | 1.75 | 2.11 | 2.09 | 2.18 | 2.05 | 1.87 | 2.19 | 2.40 | 2.94 | 2.41 | 1.87 | 1.35 | 2.42 | 2.87 | 3.99 | 2.27 | 3.98 | 2.42 | 0.77 |
| AMBER96 | 1.18 | 2.49 | 1.79 | 2.15 | 1.75 | 2.37 | 1.71 | 1.81 | 1.98 | 1.72 | 1.67 | 1.89 | 1.88 | 2.02 | 2.59 | 1.60 | 1.77 | 2.22 | 1.99 | 2.09 | 1.93 | 0.32 |
| AMBER99 | 2.47 | 1.38 | 1.47 | 1.29 | 1.24 | 1.30 | 1.81 | 1.44 | 1.30 | 1.25 | 1.90 | 2.47 | 1.72 | 1.62 | 0.60 | 1.74 | 2.87 | 3.92 | 1.23 | 2.22 | 1.76 | 0.72 |
| AMBER99SB | 2.29 | 2.32 | 2.35 | 2.59 | 2.02 | 3.16 | 1.94 | 2.24 | 2.36 | 2.25 | 2.28 | 2.70 | 2.65 | 2.42 | 2.65 | 1.91 | 2.00 | 2.57 | 2.72 | 3.02 | 2.42 | 0.33 |
| AMBER03 | 2.43 | 1.48 | 1.57 | 1.87 | 1.39 | 2.81 | 1.16 | 1.23 | 1.30 | 1.76 | 1.12 | 2.27 | 1.73 | 2.72 | 1.18 | 2.34 | 2.01 | 3.03 | 3.72 | 4.13 | 2.06 | 0.84 |
| CHARMM27 | 2.38 | 1.86 | 1.35 | 1.49 | 1.71 | 1.86 | 0.91 | 1.97 | 1.99 | 2.30 | 1.53 | 1.82 | 0.64 | 1.26 | 1.69 | 0.99 | 4.22 | 2.64 | 2.03 | 3.53 | 1.91 | 0.82 |
| OPLS-AA | 4.13 | 2.00 | 2.00 | 2.13 | 2.33 | 2.65 | 2.29 | 2.71 | 2.75 | 2.42 | 1.94 | 2.21 | 2.24 | 2.57 | 3.59 | 2.11 | 1.86 | 2.16 | 3.23 | 3.36 | 2.53 | 0.59 |
| OPLS-AA/L | 4.78 | 4.02 | 1.97 | 2.04 | 2.36 | 1.80 | 1.79 | 2.69 | 2.71 | 2.42 | 1.75 | 2.59 | 3.11 | 2.49 | 2.74 | 2.13 | 1.95 | 1.70 | 2.77 | 3.37 | 2.56 | 0.77 |
| AMBERUA | 2.56 | 1.53 | 1.22 | 1.94 | 1.01 | 1.54 | 0.92 | 1.28 | 1.91 | 1.37 | 1.59 | 2.16 | 0.96 | 0.83 | 1.30 | 1.83 | 1.44 | 2.62 | 1.36 | 3.47 | 1.64 | 0.64 |
| GROMOS(G43b1) | 2.33 | 3.31 | 3.18 | 3.14 | 3.34 | 3.07 | 2.92 | 3.14 | 3.22 | 2.95 | 2.06 | 3.00 | 2.65 | 3.15 | 3.24 | 2.71 | 7.60 | 5.63 | 3.75 | 6.22 | 3.53 | 1.33 |
| GROMOS(G45a3) | 2.47 | 3.48 | 3.38 | 3.25 | 3.54 | 3.11 | 3.09 | 3.40 | 3.43 | 3.10 | 2.25 | 3.15 | 2.83 | 3.30 | 3.40 | 2.83 | 5.24 | 5.66 | 3.62 | 5.32 | 3.49 | 0.88 |
| GROMOS(G53a6) | 7.15 | 4.68 | 4.56 | 4.34 | 4.80 | 4.22 | 3.72 | 4.60 | 4.62 | 4.25 | 3.52 | 4.17 | 3.87 | 4.24 | 4.24 | 3.85 | 4.33 | 4.71 | 4.47 | 5.13 | 4.47 | 0.72 |

[a] AMBER99 dispersion correction.

nitrogen in PM3MM ($\mu = 5.92$ kcal/mol). The additions of MM-dispersion interactions slightly improve their performance by 0.43 kcal/mol for AM1 and 0.78 kcal/mol for PM3 and PM3MM. Note that the mean RMS-C of AM1, 2.91 kcal/mol, is comparable with that (3.24 kcal/mol) of the least accurate B3LYP/cc-pVTZ method. The standard RMS-C deviations ($\sigma$(AM1) = 1.02 kcal/mol and $\sigma$(PM3) = 1.90 kcal/mol) also indicate that the examined semiempirical methods have less consistent descriptions of the five conformations than DFT methods which have $\sigma$ values ranging from 0.2−0.8 kcal/mol.

Remarkably, the polarizable force field, AMOEBA ($\mu$ = 1.56 kcal/mol), performs better than the B3LYP/cc-pVTZ ($\mu$ = 3.24 kcal/mol), B3LYP/6-31G** ($\mu$ = 1.90 kcal/mol), and semiempirical AM1 ($\mu$ = 2.91) and PM3 ($\mu$ = 5.55 kcal/mol) but less accurate than M05-2X with both cc-pVTZ or 6-31G** ($\mu \approx 0.8$ kcal/mol). The standard RMS-C deviation ($\sigma$ = 0.19 kcal/mol) of AMOEBA reaches the value of M05-2X, indicating that the force field has a consistent description over all five conformations. These are encouraging signs for developing such a polarizable force field to simulate biological molecules. However, the performance of the two versions

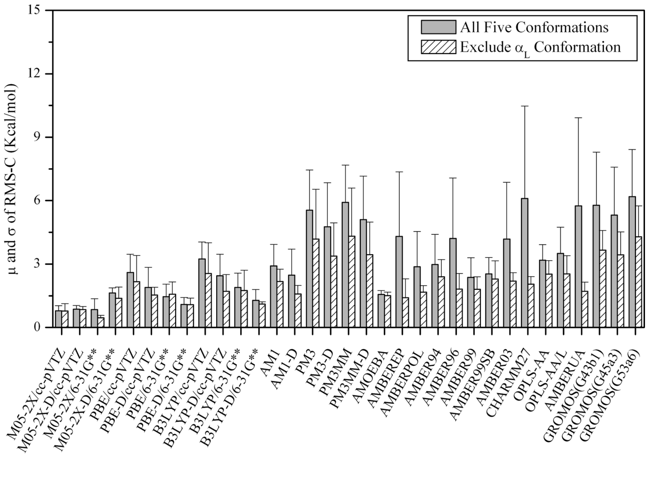


***Figure 3.*** Mean ($\mu$) and standard deviation ($\sigma$) of RMS-C of each conformation calculated over 20 amino acids for each method.

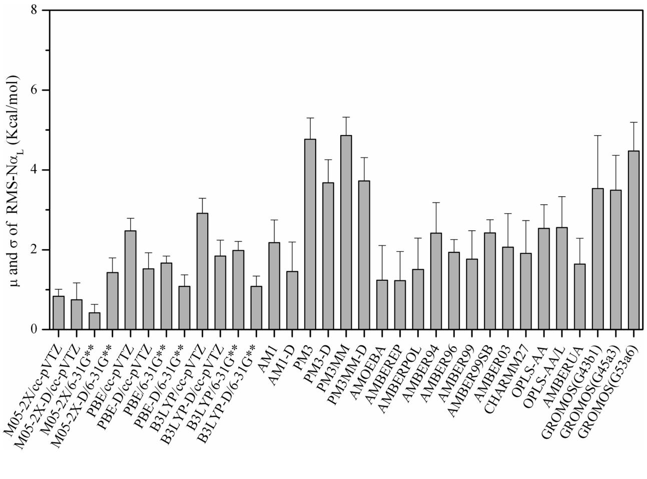


***Figure 4.*** Mean ($\mu$) and standard deviation ($\sigma$) of RMS of each amino acid calculated without $\alpha_L$ conformations for each method.

of AMBER polarizable force fields (with/without extra points) is not improved in comparison with their nonpolarizable AMBER versions (see Table 1). A systematic parametrization is necessary to improve the accuracy.

Unexpectedly, some of the additive all-atom force fields outperform semiempirical methods (see Table 2 for comparing the mean RMS-C of these force fields with the semiempirical methods). Large standard deviations of RMS-C are observed for AMBER96 (2.86 kcal/mol), AMBER03 (2.69 kcal/mol), and CHARMM27 (4.37 kcal/mol), indicating that these force fields have imbalanced descriptions on some conformations. Indeed, these three force fields have larger RMS-C's for the $\alpha_L$ conformation, being 9.55, 9.11, and 14.55 kcal/mol, respectively, than the other four conformations. This defect is probably caused by the less attention paid to the $\alpha_L$ conformation in the force field parametrization. The OPLS-AA/OPLS-AA/L and AMBER99/AMBER99SB force fields suffer such a defect less severely; their $\sigma$ values are 0.74/1.23 and 0.93/0.78 kcal/mol, respectively. If excluding the $\alpha_L$ conformation, the mean RMS-C-N$\alpha_L$ of AMBER96/99 and CHARMM27 are respectively 1.82/1.81 and 2.05 kcal/mol (Table S4c of SIA), which are smaller than the best semiempirical AM1 method (2.18 kcal/mol of AM1) but are still larger than the polarizable AMOEBA force field (1.50 kcal/mol). We emphasized that the tetrapeptide models used in this study are no longer the alanine dipeptide which is often used in force field parametrization, and they can mimic the H-bonds in protein helix secondary structures. For the AMBER series, the AMBER99 and AMBER96 perform slightly better than the others in the gas phase after excluding the $\alpha_L$ data. However, caution should be taken that AMBER99SB and AMBER03 were developed to implicitly include the solvent effect. Further evaluation in the condensed phase is necessary. For the OPLS series of force fields, it is unexpected that OPLS-AA/L performs slightly worse than OPLS-AA, because the torsion parameters of the former were refined using QM-based conformational energies of a large amount of different conformations in the gas phase. Again, further evaluation in the condensed phase is necessary to provide a more reasonable assessment.[18]

In the category of the united atom force fields, AMBERUA is comparable to the GROMOS96 series when all five conformations are included. But when $\alpha_L$ is excluded, AMBERUA performs much better than the GROMOS series (~1.7 kcal/mol vs ~4.0 kcal/mol of mean RMS-C). Since the versions of G45a3 and G53ab were optimized in the solution phase, further evaluation in the solution phase is necessary. In addition, we argue that other important aspects are needed to be considered for a comprehensive assessment of the force fields (especially for the additive ones): interpeptide interactions,[66,79,80] peptide−water interactions,[14,71] thermodynamic properties,[15,16] even kinetic properties, etc. In fact, the GROMOS force fields, which perform worse than others in this study, have been successfully applied to many protein simulations.

The means ($\mu$) of RMS (Table S3d, Supporting Information) and RMS-N$\alpha_L$ (Table S4d/Table 3) give the same information on the overall performance of the examined methods as those of RMS-C (Table S3c/Table 2) and RMS-C-N$\alpha_L$ (Table S4c) because they originate from the same data sets. However, the individual RMS/RMS-N$\alpha_L$ can tell us the performance of the examined methods on the individual amino acid. It is well-known and we also confirmed that the MM methods are not able to describe the $\alpha_L$ conformation properly. We thus use RMS-N$\alpha_L$ (Table 3 and Figure 4) for the following discussion, since $\alpha_L$ is not important for modeling the native protein structure. Understandably, the QM methods (M05-2X, PBE, B3LYP, AM1, and PM3) are generally more consistent in describing all 20 amino acids, although some of them (e.g., PBE/cc-pVTZ, B3LYP/cc-pVTZ, PM3, and PM3MM) have poor overall performance. The polarizable AMOEBA, which has overall good performance, is unsatisfied with some amino acids such as Pro, Asp, and Glu in particular. The developer needs to pay attention to these problematic amino acids. For the additive force fields, AMBER96 gives more consistent descriptions to all the amino acids than the other additive force fields. Its standard deviation, 0.32 kcal/mol, is comparable with those of the M05-2X methods, although the overall performance of the force field is not as good as those of the M05-2X methods. The readers can refer the Table 3 to identify the problematic amino acids for other force fields.

## 4. Conclusions

Using 100 tetrapeptide structures which cover all 20 amino acids and five major conformations ($\alpha_R$, $\alpha_L$, $\beta$, $\beta_a$, and PPII), we estimated their conformation energies in the gas phase at the MP2/cc-pVTZ//B3LYP/6-31G** level. The results indicate that the energetic patterns (the order and the energy gap) of the five conformations bear certain resemblances among the amino acids in the same class but is quite different among the amino acids in the different classes (e.g., hydrophobic, aromatic, polar, and charged classes). Using these MP2 energies of 100 tetrapeptide structures as "standard", we further evaluated the performance of various methods in terms of RMS and RMS-C and draw the following conclusions: (1) The M05-2X DFT functional outperforms PBE and B3LYP. (2) The semiempirical methods (AM1, PM3, and PM3MM) are not accurate enough to describe the relative energies of the conformations. (3) The AMOEBA polarizable force field outperforms the semiempirical methods and the B3LYP method. However, the current AMBER polarizable force fields do not improve the accuracy with respect to the related additive versions, which suggest a systematic parametrization is necessary to improve the accuracy. (4) The additive force fields are less accurate than the three DFT methods, but some of them are more accurate than the semiempirical methods. (5) If excluding the $\alpha_L$ conformation, the examined force fields have comparable performance; the RMS-C means are 2.4 kcal/mol for AMBER94, 1.8 kcal/mol for AMBER96/99, 2.3 kcal/mol for AMBER99SB, 2.2 kcal/mol for AMBER03, 2.0 kcal/mol for CHARMM27, and 2.5 kcal/mol for OPLS-AA and OPLS-AA/L. However, it should be pointed out that some of the force fields are parametrized to include the solvent effects implicitly, while our calculations were carried out in the gas phase. (6) If excluding the $\alpha_L$ conformation, the united atom force field (AMBERUA) with a 1.7 kcal/mol of mean RMS-C, has an accuracy comparable with that of the all-atom force field. (7) With respect to the MP2 energies, overestimating the energies of the compact helical conformations ($\alpha_R$ and $\alpha_L$), but underestimating those of the extended conformations ($\beta$, $\beta_a$, and PPII), is a general error trend for methods M05-2X/cc-pVTZ, PBE and B3LYP, AM1, AMBER99SB, OPLS-AA, and OPLS-AA/L and GROMOS. (8) Semiempirical and empirical force field methods perform poorly on Pro and the charged amino acids.

The structures and energies of the 100 tetrapeptide structures can serve as a database to systematically develop/calibrate force fields for modeling proteins. In addition to the data provided in the Supporting Information, other preliminary data such as the Cartesian coordinates of the 100 tetrapeptides are available upon request.

**Acknowledgment.** This study was supported by the Chinese Academy of Sciences, NSFC (No: 20773160), and the Ministry of Science and Technology (No: 2009ZX09501-011). We also express our gratitude to the reviewers for their comments and insights which targeted revision of the paper.

**Supporting Information Available:** The main-chain and side-chain dihedral angles adopted in the calculations (Table S1 of SIA). The relative energies of $\alpha_L$, $\beta$, $\beta_a$, and PPII to $\alpha_R$ at all considered levels (Table S2 of SIA and Table S1 of SIB). The reference energy offsets ($E_C$/$E_C$ − N$\alpha_L$) to minimize RMS/RMS-N$\alpha_L$ (Table S3a/Table S4a of SIA and Table S2a/Table S3a of SIB). The signed energy errors (error = $E_{ai} - E_{bi} + E_c$) and errors (Table S3b of SIA and Table S2b of SIB)/error-N$\alpha_L$ (Table S4b of SIA and Table S3b of SIB). The signed RMS-C/RMS-C-N$\alpha_L$ values (Table S3c/Table S4c of SIA and Table S2c and Table S3c of SIB). The unsigned RMS/RMS-N$\alpha_L$ values (Table S3d/Table S4d of SIA and Table S2d/Table S3d of SIB). Representative geometries for five conformations of each tetrapeptide (Figure S1 of SIA). Relative energies of the MP2/cc-pVTZ//M05-2X/6-31G** method (Figure S1 of SIB). The signed energy errors, signed RMS-C/RMS-C-N$\alpha_L$ values, and RMS values of each method are plotted in Figure S2-S34 of SIA and Figure S2-S34 of SIB. Mean and stand deviation of RMS-C of each conformation calculated over 20 amino acids relative to MP2/cc-pVTZ//M05-2X/6-31G** method (Figure S35 of SIB). Comparison of the RMS-N$\alpha_L$ using two set of geometries (B3LYP/6-31G** and M05-2X/6-31G**; Figure S36 of SIB). This material is available free of charge via the Internet at http://pubs.acs.org.

## References

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

(2) Moller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618–622.

(3) Axel, D. B. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(4) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623–11627.

(5) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(6) Yan, Z.; Donald, G. T. *J. Chem. Phys.* **2006**, *125*, 194101–194118.

(7) Hohenstein, E. G.; Chill, S. T.; Sherrill, C. D. *J. Chem. Theory Comput.* **2008**, *4*, 1996–2000.

(8) Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 289–300.

(9) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.

(10) James, J. P. S. *J. Comput. Chem.* **1989**, *10*, 209–220.

(11) James, J. P. S. *J. Comput. Chem.* **1989**, *10*, 221–264.

(12) Jenn-Huei, L.; Norman, L. A. *J. Comput. Chem.* **1991**, *12*, 186–199.

(13) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.

(14) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(15) Schuler, L. D.; Daura, X.; Gunsteren, W. F. v. *J. Comput. Chem.* **2001**, *22*, 1205–1218.

(16) Oostenbrink, C.; Villa, A.; Mark, A. E.; Gunsteren, W. F. V. *J. Comput. Chem.* **2004**, *25*, 1656–1676.

(17) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

(18) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.

(19) Wang, Z.-X.; Zhang, W.; Wu, C.; Lei, H.; Cieplak, P.; Duan, Y. *J. Comput. Chem.* **2006**, *27*, 781–790.

(20) Dixon, R. W.; Kollman, P. A. *J. Comput. Chem.* **1997**, *18*, 1632–1646.

(21) Guillaume, L.; Benoit, R. *J. Chem. Phys.* **2003**, *119*, 3025–3039.

(22) Lamoureux, G.; MacKerell, A. D.; Roux, B. *J. Chem. Phys.* **2003**, *119*, 5185–5197.

(23) Geerke, D. P.; van Gunsteren, W. F. *J. Chem. Theory Comput.* **2007**, *3*, 2128–2137.

(24) Jorgensen, W. L.; Jensen, K. P.; Alexandrova, A. N. *J. Chem. Theory Comput.* **2007**, *3*, 1987–1992.

(25) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A. *J. Phys. Chem. A* **2003**, *108*, 621–627.

(26) Friesner, R. A.; Robert, L. B.; David, B. *Adv. Protein Chem.* **2005**, *72*, 79–104.

(27) Yang, Z.-Z.; Wang, C.-S. *J. Phys. Chem. A* **1997**, *101*, 6315–6321.

(28) Ponder, J. W.; Case, D. A.; Valerie, D. *Adv. Protein Chem.* **2003**, *66*, 27–85.

(29) Nohad, G.; Sherif, A. K.; Jean-Francois, T.; Dennis, R. S. *J. Comput. Chem.* **2004**, *25*, 823–834.

(30) Wang, J.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049–1074.

(31) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999–2012.

(32) Head-Gordon, T.; Head-Gordon, M.; Frisch, M. J.; Brooks, C. L.; Pople, J. A. *J. Am. Chem. Soc.* **1991**, *113*, 5989–5997.

(33) Viviani, W.; Rivail, J. L.; Perczel, A.; Csizmadia, I. G. *J. Am. Chem. Soc.* **1993**, *115*, 8321–8329.

(34) Alemán, C.; Casanovas, J. *J. Chem. Soc., Perkin Trans. 2* **1994**, 563–568.

(35) Gould, I. R.; Cornell, W. D.; Hillier, I. H. *J. Am. Chem. Soc.* **1994**, *116*, 9250–9256.

(36) Gronert, S.; O'Hair, R. A. J. *J. Am. Chem. Soc.* **1995**, *117*, 2071–2081.

(37) Cornell, W. D.; Gould, I. R.; Kollman, P. A. *THEOCHEM* **1997**, *392*, 101–109.

(38) Mohle, K.; Gussmann, M.; Rost, A.; Cimiraglia, R.; Hofmann, H. J. *J. Phys. Chem. A* **1997**, *101*, 8571–8574.

(39) Kaschner, R.; Hohl, D. *J. Phys. Chem. A* **1998**, *102*, 5111–5116.

(40) Aleman, C. *J. Phys. Chem. A* **2000**, *104*, 7612–7616.

(41) Barroso, M. N.; Cerutti, E. S.; Rodríguez, A. M.; Jáuregui, E. A.; Farkas, O.; Perczel, A.; Enriz, R. D. *THEOCHEM* **2001**, *548*, 21–37.

(42) Masman, M. F.; Amaya, M. G.; Rodríguez, A. M.; Suvire, F. D.; Chasse, G. A.; Farkas, O.; Perczel, A.; Enriz, R. D. *THEOCHEM* **2001**, *543*, 203–222.

(43) Yu, C.-H.; Norman, M. A.; Schafer, L.; Ramek, M.; Peeters, A.; van Alsenoy, C. *J. Mol. Struct.* **2001**, *567−568*, 361–374.

(44) Aleman, C.; Jimenez, A. I.; Cativiela, C.; Perez, J. J.; Casanovas, J. *J. Phys. Chem. B* **2002**, *106*, 11849–11858.

(45) Vargas, R.; Garza, J.; Hay, B. P.; Dixon, D. A. *J. Phys. Chem. A* **2002**, *106*, 3213–3218.

(46) Andras, P.; Odon, F.; Imre, J.; Igor, A. T.; Imre, G. C. *J. Comput. Chem.* **2003**, *24*, 1026–1042.

(47) Klipfel, M. W.; Zamora, M. A.; Rodriguez, A. M.; Fidanza, N. G.; Enriz, R. D.; Csizmadia, I. G. *J. Phys. Chem. A* **2003**, *107*, 5079–5091.

(48) Chin, W.; Mons, M.; Dognon, J. P.; Mirasol, R.; Chass, G.; Dimicoli, I.; Piuzzi, F.; Butz, P.; Tardivel, B.; Compagnon, I.; vonHelden, G.; Meijer, G. *J. Phys. Chem. A* **2005**, *109*, 5281–5288.

(49) Masman, M. F.; Lovas, S.; Murphy, R. F.; Enriz, R. D.; Rodriguez, A. M. *J. Phys. Chem. A* **2007**, *111*, 10682–10691.

(50) Young Kee, K.; Nam Sook, K. *J. Comput. Chem.* **2009**, *30*, 1116–1127.

(51) Möhle, K.; Höfmann, H.-J. *J. Mol. Model.* **1998**, *4*, 53–60.

(52) Langella, E.; Rega, N.; Improta, R.; Crescenzi, O.; Barone, V. *J. Comput. Chem.* **2002**, *23*, 650–661.

(53) Schlund, S.; Müller, R.; Gra*β*mann, C.; Engels, B. *J. Comput. Chem.* **2008**, *29*, 407–415.

(54) Böehm, H. J.; Brode, S. *J. Am. Chem. Soc.* **1991**, *113*, 7129–7135.

(55) Gould, I. R.; Kollman, P. A. *J. Phys. Chem.* **1992**, *96*, 9255–9258.

(56) Beachy, M. D.; Chasman, D.; Murphy, R. B.; Halgren, T. A.; Friesner, R. A. *J. Am. Chem. Soc.* **1997**, *119*, 5908–5920.

(57) Kaminsky, J.; Jensen, F. *J. Chem. Theory Comput.* **2007**, *3*, 1774–1788.

(58) Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1988**, *110*, 1657–1666.

(59) Improta, R.; Barone, V. *J. Chem. Phys.* **2001**, *114*, 2541–2549.

(60) Improta, R.; Barone, V.; Kudin, K. N.; Scuseria, G. E. *J. Am. Chem. Soc.* **2001**, *123*, 3311–3322.

(61) Ireta, J.; Neugebauer, J.; Scheffler, M.; Rojo, A.; Galvan, M. *J. Phys. Chem. B* **2003**, *107*, 1432–1437.

(62) Bogár, F.; S., Z.; Bartha, F.; Penke, B.; Ladik, J. *Phys. Chem. Chem. Phys.* **2005**, *7*, 2965–2969.

(63) Ireta, J.; Neugebauer, J.; Scheffler, M.; Rojo, A.; Galvan, M. *J. Am. Chem. Soc.* **2005**, *127*, 17241–17244.

(64) Ireta, J.; Scheffler, M. *J. Chem. Phys.* **2009**, *131*, 085104−085108.

(65) Penev, E.; Ireta, J.; Shea, J.-E. *J. Phys. Chem. B* **2008**, *112*, 6872–6877.

(66) Mackerell, A. D.; Feig, M.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 1400–1415.

(67) MacKerell, A. D.; Feig, M.; Brooks, C. L. *J. Am. Chem. Soc.* **2004**, *126*, 698–699.

(68) Yoda, T.; Sugita, Y.; Okamoto, Y. *Chem. Phys.* **2004**, *307*, 269–283.

(69) Shell, M. S.; Ritterson, R.; Dill, K. A. *J. Phys. Chem. B* **2008**, *112*, 6878–6886.

(70) Dunbrack, R. L. *Curr. Opin. Struct. Biol.* **2002**, *12*, 431–440.

(71) Wang, Z.-X.; Duan, Y. *J. Comput. Chem.* **2004**, *25*, 1699–1716.

(72) 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.; N. Rega; 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.

(73) Ponder, J. W. *Tinker*, version 4.2; Biochemistry & Molecular Biophysics, Washington University School of Medicine: St. Louis, MO, 2004.

(74) Kollman, P.; Dixon, R.; Cornell, W.; Fox, T.; Chipot, C.; Pohorille, A. In *Computer Simulation of Biomolecular Systems*; Wilkinson, A.,Weiner, P., van Gunsteren, W. F., Eds.; Elsevier: New York, 1997; Vol. 3, pp 83−96.

(75) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 712–725.

(76) Yang, L.; Tan, C.-h.; Hsieh, M.-J.; Wang, J.; Duan, Y.; Cieplak, P.; Caldwell, J.; Kollman, P. A.; Luo, R. *J. Phys. Chem. B* **2006**, *110*, 13166–13176.

(77) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; Schafmeister, C.; Ross; W. S. Kollman, P. A. *Amber 9*; University of California: San Francisco, 2006.

(78) Spoel, D. V. D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701–1718.

(79) Alexander, D.; Mackerell, M., Jr. *J. Comput. Chem.* **2004**, *25*, 1400–1415.

(80) Wang, Z.-X.; Wu, C.; Lei, H.; Duan, Y. *J. Chem. Theory Comput.* **2007**, *3*, 1527–1537.

CT100008Q

JCTC Journal of Chemical Theory and Computation

# Coarse-Grained Model of Collagen Molecules Using an Extended MARTINI Force Field

Alfonso Gautieri,[†,‡] Antonio Russo,[†] Simone Vesentini,[†] Alberto Redaelli,[†] and Markus J. Buehler*[,‡,§]

*Biomechanics Group, Department of Bioengineering, Politecnico di Milano, Via Golgi 39, 20133 Milan, Italy, Laboratory for Atomistic and Molecular Mechanics, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, 77 Massachusetts Ave., Room 1-235A&B, Cambridge, Massachusetts, and Center for Computational Engineering, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, Massachusetts*

Received January 9, 2010

**Abstract:** Collagen is the most abundant protein in the human body, providing mechanical stability, elasticity, and strength to connective tissues such as tendons, ligaments, and bone. Here, we report an extension of the MARTINI coarse-grained force field, originally developed for lipids, proteins, and carbohydrates, used to describe the structural and mechanical properties of collagen molecules. We identify MARTINI force field parameters to describe hydroxyproline amino acid residues and for the triple helical conformational structure found in collagen. We validate the extended MARTINI model through direct molecular dynamics simulations of Young's modulus of a short 8-nm-long collagen-like molecule, resulting in a value of approximately 4 GPa, in good agreement with earlier full atomistic simulations in explicit solvent as well as experimental results. We also apply the extended MARTINI model to simulate a 300-nm-long human type I collagen molecule with the actual amino acid sequence and calculate its persistence length from molecular dynamics trajectories. We obtain a value of 51.5 ± 6.7 nm for the persistence length, which is within the range of earlier experimental results. Our work extends the applicability of molecular models of collagenous tissues by providing a modeling tool to study collagen molecules and fibrils at much larger scales than accessible to existing full atomistic models, while incorporating key chemical and mechanical features and thereby presenting a powerful approach to computational materiomics.

## 1. Introduction

Collagen (or tropocollagen) molecules represent the most abundant protein building block in the human body, where they provide primarily mechanical stability, elasticity, and strength to connective tissues such as tendons, ligaments, and bone.[1] Collagen's primary structure is formed by a sequence of triplets (glycine−X−Y)$_n$, where X and Y can represent any amino acids, but mostly proline and hydroxyproline. In particular, the presence of glycine every three amino acids guarantees the stability of the tertiary molecular structure, characterized by a right-handed triple helix[2] with an overall length of ∼300 nm and diameter of ∼1.5 nm, leading to a great aspect ratio of ∼200.

Collections of collagen molecules form well-defined hierarchical structures in tissues that give rise to fibrils and fibers, which universally represent the basis of most connective tissues.[3] Collagen provides mechanical integrity to connective tissues through, above all, a great resistance to

* Corresponding author phone: +1-617-452-2750; e-mail mbuehler@MIT.EDU.

† Politecnico di Milano.

‡ Department of Civil and Environmental Engineering, Massachusetts Institute of Technology.

§ Center for Computational Engineering, Massachusetts Institute of Technology.

10.1021/ct100015v
Published on Web 03/24/2010

stretching along the main direction of molecules, fibrils, and fibers. Moreover, collagen is capable of providing outstanding mechanical features as it interacts with other biological molecules and minerals, such as glycosaminoglycans or hydroxyapatite in bone. For example, in bone tissue, the toughness is much greater than that of collagen or hydroxyapatite alone. This is due to the fact that collagen, with its capacity to yield and dissipate energy, contributes to the material's enhanced toughness by providing an effective means to confer shear stresses between hydroxyapapite crystals and to dissipate mechanical energy through molecular unfolding and intermolecular shear.[4] The previous examples underline that, physiologically, collagen cannot solely be considered as a single molecule but must be studied in the context of functional networks that are formed on the basis of a large number of collagen molecules and other biopolymers. The significance of collagen in controlling key material properties in connective tissues is further evident from mutations in collagen, which result in incorrect protein folding that causes several severe pathologies, such as Ehlers−Danlos syndrome, Alport syndrome,[5] or *osteogenesis imperfecta*.[6]

Great efforts have been put forth in recent years focused on characterizing the mechanical properties of collagen molecules and fibrils using both experimental and computational (theoretical) approaches. Some of the earliest experiments analyzed type I collagen molecules in terms of their stiffness. For example, Sasaki and Odajima[7] estimated the Young's modulus of collagen molecules by using X-ray diffraction techniques. Bozec and Horton[8] used an AFM setup to evaluate both the topographical and mechanical properties of the collagen molecules. Optical trapping methods have also been used for the assessment of collagen's mechanical properties, as reported by Sun and co-workers.[9] A review of the global efforts in the understanding of the structure and properties of collagen-based tissues are nicely illustrated in recent works.[10] Nonetheless, a persistent limitation of experimental analyses is the lack of details about nanoscale phenomena, and limitations in sample preparation. To complement experimental approaches, molecular simulation provides an alternative approach to describe the molecular mechanics of collagen from a bottom-up perspective. Most earlier studies, often based on collagen-like peptides obtained from X-ray crystallography,[11] focused on relatively short collagen molecules[12–17] limited to ∼10 nm length. However, the contour length of collagen molecules is ∼300 nm, which represents a scale that is not yet accessible to full atomistic molecular modeling. Further, the study of collagen fibrils and fibers is currently not feasible with full atomistic simulation, which represents an important limitation since this scale in the structural hierarchy of collagen is most relevant for physiological function. Other limitations exist with respect to the accessible time scale, where most full-atomistic simulations are limited by a few hundred nanoseconds. However, many relevant materials phenomena such as tissue deformation and failure (e.g., under disease conditions) emerge at much longer time scales.

A promising strategy to overcome these limitations is to decrease the number of degrees of freedom by grouping atoms into pseudoatoms (or particles) referred to as beads. This represents the basis of the so-called coarse-grained approach,[18,19] where, starting at the nanoscale, it is possible to derive parameters for higher hierarchical levels, up to the macroscale by systematically feeding information from smaller, more accurate to larger, more coarse levels. The development of a coarse-grained model provides a powerful path toward reliable modeling of full-length collagen molecules and its higher-level hierarchical structures. Specifically, coarse-grained models allow the study of more complex systems, up to micrometer dimension and millisecond duration.[20] Earlier work of coarse-graining collagen molecules grouped hundreds of atoms into particles or beads.[14,15] This level of coarse-graining, however, is at a relatively coarse level where information about biochemical features (e.g., amino acid sequence) cannot be represented directly. However, the incorporation of biochemical features is crucial in a computational materiomics approach where material properties are elucidated at multiple scales, including the level of amino acid sequence that links to genetic level information.

Several other coarse-grained models suitable for proteins have been developed and successfully applied in earlier works.[21] Particularly, the MARTINI coarse-grained model,[22] developed by Marrink and co-workers, was initially applied to membrane lipids, later extended to proteins,[23] and recently also to carbohydrates.[24] It has been successfully applied to gain insights into different biological molecules such as membrane proteins,[25,26] ion channels,[27] and liposomes.[28]

The MARTINI model provides a suitable level of coarse-graining, as it retains information about the chemistry specific to the amino acid sequence (as side chains are modeled depending on the type of the amino acids). All amino acids in the MARTINI force field are modeled with a number of beads that varies depending on the steric volume of each amino acid.[23] The general mapping rule is that four heavy atoms (that is, non-hydrogen atoms) are grouped together into one bead. Further, one bead describes the backbone, while others are added to represent the side chain. The number of beads used to model a specific residue therefore varies depending on the dimensions of the side chain of the amino acid. Small amino acids such as glycine or alanine are described by just one bead, while larger amino acids, like phenylalanine, tyrosine, and tryptophan are modeled with up to five beads. The model also takes in account the polarity of every bead, described by a letter (P, polar; C, apolar; N, nonpolar; Q, charged) and a number (from 1, low polarity, to 5, high polarity). Further, a letter is used to characterize a residue's hydrogen bonding capability (d, donor; a, acceptor; da, donor and acceptor; 0, none). These bead types, which correspond to atom types in the atomistic modeling framework, are used to describe the nonbonded interactions between beads. For each pair of bead types, a set of parameters for electrostatic and van der Waals potentials is defined. The best choice of particle types for amino acids was obtained by the authors of the original MARTINI force field for proteins by comparing simulation results and experimental measurements of water/oil partitioning coefficients of the amino acid side-chain analogues.[23]

However, although the MARTINI force field is suitable for a wide number of applications and its validity has been proven through several simulations and comparisons both with atomistic and with experimental data, it cannot be applied to model collagen molecules in the presently available formulation. This is because of two reasons. First, the existing formulations of MARTINI lack parameters for hydroxyproline (a nonstandard amino acid, found solely in collagen and formed through post-translational modification of proline). Second, it lacks parameters to describe the triple helical configuration of a collagen molecule. The work reported here overcomes these limitations and addresses two major points. First, we extend the MARTINI model to enable the modeling of collagen molecules by including parameters for hydroxyproline and by adding parameters to describe the collagen triple helix. Second, the extended MARTINI force field is applied to perform coarse-grained molecular dynamics simulations of a full-length model of a collagen type I molecule, used to calculate mechanical parameters such as Young's modulus and the persistence length, which allows us to validate the new model through a comparison with earlier computational and experimental results.

## 2. Materials and Methods

We use a computational multiscale approach to generate an atomistic-informed coarse-grained model of tropocollagen in the framework of the MARTINI force field. The parameters of the coarse-grained model are obtained through a combination of experimental and full-atomistic modeling data to identify the parameters for hydroxyproline residues, as well as through a statistical analysis of collagen-like PDB entries to assess the geometrical features of the coarse-grained tropocollagen triple helix. The results of the parametrization are fed into the coarse-grained tropocollagen model in the spirit of a multiscale simulation approach.

All full atomistic simulations are carried out using the GROMACS code[29,30] and the GROMOS96 43a1 force field, which includes also parameters for the hydroxyproline residue. The protein molecules are entirely solvated in periodic water boxes with single point charge (SPC) water molecules as the solvent model. SETTLE (for water) and LINCS algorithms are used to constrain covalent bond lengths involving hydrogen atoms, thus allowing a time step of 2 fs. Nonbonding interactions are computed using a cutoff for the neighbor list at 1 nm, with a switching function between 0.8 and 0.9 nm for van der Waals interactions, while the Particle-Mesh Ewald sums (PME) method is applied to describe electrostatic interactions. In the case of charged peptides, counterions ($Cl^-$ or $Na^+$) are added in order to keep the system neutral. The preliminary system energy minimizations are performed by using a steepest descent algorithm until convergence. The systems are then equilibrated through *NPT* molecular dynamics simulations at a temperature of 310 K (37 °C). In order to assess the elastic constants of the coarse-grained model, simple atomistic oligopeptides are considered. A pulling force is applied along the molecular axis, where one molecular end is held fixed using a rigid constraint while the other is pulled through the use of a virtual spring with a known elastic constant. This setup corresponds to the steered molecular dynamics setup used in similar atomistic works.[13,31] In particular, the value of the spring elastic constant is chosen to be 4000 kJ $mol^{-1}$ $nm^{-2}$, while the deformation rate is set at 0.1 m/s. The steered molecular dynamics simulations performed on the atomistic systems provide the reference force-extension behaviors.

Coarse-grained molecular dynamics simulations are carried out using the GROMACS code and the MARTINI force filed with the inclusion of the parameters found in this work for tropocollagen. The models are entirely solvated in periodic water boxes using the coarse-grained water model provided within the MARTNI force field. In the case of charged molecules, counterions ($Cl^-$ or $Na^+$) are added in order to keep the system neutral. The preliminary system energy minimizations are performed by using a steepest descent algorithm until convergence. The systems are then equilibrated through NPT molecular dynamics simulations at a temperature of 310 K (37 °C) using a time step of 20 fs. Steered molecular dynamics simulations are performed using the same setup as described for atomistic simulations. Finally, the whole structure of the heterotrimeric type I collagen molecule is studied. The primary structure is obtained from PubMed (entry number NP_000079 for $\alpha_1$ chain and NP_000080 for $\alpha_2$ chain), and the atomistic triple helical structure is built using the software THeBuScr[32] (Triple-Helical collagen Building Script). The atomistic representation is then coarse-grained and divided into five 60-nm-long segments, which are energy minimized and simulated in explicit solvent at finite temperature for 350 ns each.

## 3. Results

**3.1. Extension of the MARTINI Force Field to Include Hydroxyproline.** The existing MARTINI protein force field represents all 20 naturally occurring amino acids but lacks hydroxyproline, an amino acid produced in collagen synthesis via post-translational modification of proline (Figure 1A). In order to extend the MARTINI model for the study of collagen molecules, hydroxyproline parameters must be introduced. In the MARTINI model, proline is modeled through the use of two beads, one for the backbone (bead type C5) and one for the side-chain (bead type AC2), as shown in Figure 1A. Hydroxyproline derives directly from proline via the addition of a hydroxyl group on its side-chain. Therefore, hydroxyproline is also modeled using two beads in the extended MARTINI model. For the backbone bead, we maintain the same bead type as for proline. The side chain, however, due to the presence of the hydroxyl group, shows a higher polarity level than proline. This aspect was already demonstrated in the work of Black and Mould,[33] who calculated an index that takes into account the hydrophobicity of all amino acid side-chains, including hydroxyproline. Matching the bead types assigned by Marrink and the hydrophobicity values derived from Black and Mould, the hydroxyproline side-chain bead polarity, in the MARTINI notation, is determined. This parameter ranges from 0 to 1, where 1 is the most hydrophobic amino acid (phenylalanine), while 0 is the most hydrophilic one (arginine). The value reported for hydroxyproline is 0.527, which is found between

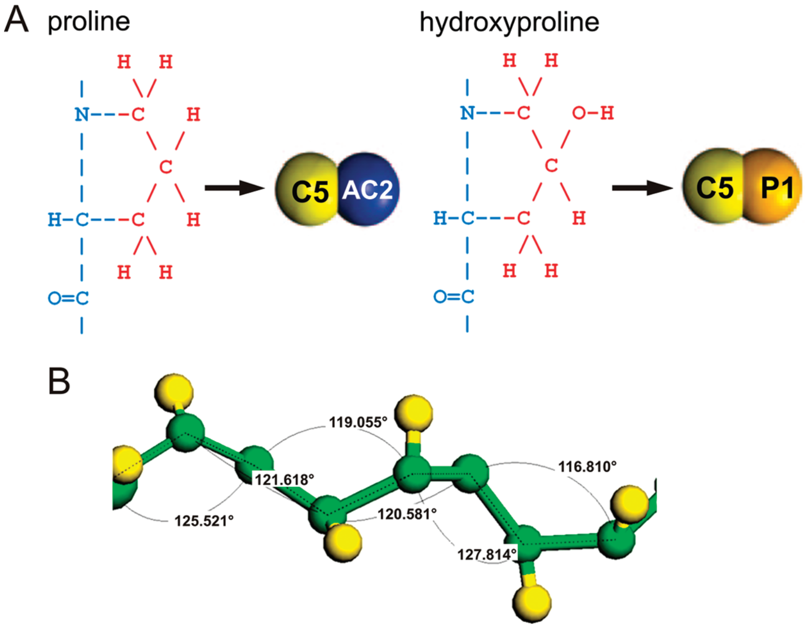


***Figure 1.*** Parameterization of the force field of the MARTINI force field including the hydroxyproline residue (panel A) and geometrical features (angles) of the coarse-grained collagen triple helix (panel B). Hydroxyproline directly derives from proline, adding a hydroxil group to its side chain, giving rise to an increased polarity level. Atomistic structures of both proline and hydroxyproline are shown on the left side of panel A, while the bead representation using MARTINI model notation is shown on the right side of panel A. Panel B shows the analysis of the angle between backbone beads of collagen-like peptides. The statistical analysis is performed on five collagen PDB entry (2DRT, 1K6F, 1QSU, 1CAG, 1V6Q). These collagen-like peptides are coarse-grained, and the bond lengths, angles, and dihedrals between backbone beads are analyzed.

the threonine (0.450) and cysteine (0.680) values. In the MARTINI force field, the threonine side chain is modeled as a P1 bead, while cysteine is modeled as a C5 bead. Considering this result, hydroxyproline is assigned a P1 value for the hydroxyproline side-chain, since its hydrophobicity value is closer to the value for threonine.

**3.2. Extension of the MARTINI Force Field to Describe the Triple Helical Structure of Collagen Molecules.** The MARTINI model also takes into account the secondary structure of a protein, such as the α-helix, $\beta$-sheet, extended turns, or bends. The characteristic triple helical structure of collagen molecules cannot be described by any of the secondary structure parameters currently included in the MARTINI model. Therefore, it is necessary to determine additional parameters for the triple helical collagen structure. To achieve this, we assess the bond distances, angles, and dihedral parameters specific to the collagen triple helix. The MARTINI bond potential forms are maintained in our extended formulations, which are described as follows:

$$V_b = \frac{1}{2}K_b(d_{ij} - d_b)^2 \qquad (1)$$

$$V_a = \frac{1}{2}K_a[\cos(\varphi_{ijk}) - \cos(\varphi_a)]^2 \qquad (2)$$

$$V_d = K_d[1 + \cos(n\psi_{ijkl} - \psi_d)] \qquad (3)$$

where $V_b$, $V_a$, and $V_d$ are the potential energy terms of bond stretching, angle deformation, and dihedral deformations, respectively. Bonded interactions act between bonded sites $i$, $j$, $k$, and $l$ with a distance at the equilibrium of $d_b$, angle $\varphi_a$, and dihedral angles $\Psi_d$ and with elastic stiffness of $K_b$, $K_a$, and $K_d$ (for bond, angle, and dihedral, respectively).

The triple helical geometrical parameters of the collagen triple helix (that is, $d_b$, $\varphi_a$, and $\Psi_d$) are obtained by performing a statistical analysis on a set of five collagen-like molecules available in the Protein Data Bank[34] (identification codes: 2DRT, 1K6F, 1QSU, 1CAG, and 1V6Q). The bond lengths between backbone beads, bonding angles, and dihedral angles are computed on the basis of these five crystallographic structures. From a statistical analysis of different collagen molecules, we determine a bond reference length ($d_b$) of 0.365 ± 0.07 nm, a bonding reference angle ($\varphi_a$) of 119.2 ± 8.72°, and a dihedral reference angle ($\Psi_d$) of $-89.3 \pm 9.76°$. Figure 1B shows the details of a coarse-grained model of a collagen chain with the angular geometrical parameters.

In order to assess the stiffnesses of bonds, angles, and dihedrals (respectively, $K_b$, $K_a$, and $K_d$), simple atomistic oligopeptides are considered, as shown in Figure 2A. In order to derive the bonding constant $K_b$, a glycine−proline structure is chosen. In order to derive the angle elastic constant $K_a$, a glycine−proline−hydroxyproline oligopeptide is considered. Finally, to obtain the dihedral elastic constants $K_d$, a glycine−proline−hydroxyproline−glycine molecule is studied. These structures are considered both in their atomistic form and in their coarse-grained form, and the parameters of the coarse-grain model (i.e., $K_b$, $K_a$, and $K_d$) are identified matching force-extension curves obtained from atomistic and coarse-grained simulations.

Steered molecular dynamics simulations are performed on the three atomistic systems to obtain reference force-extension behaviors. For the coarse-grained models, the

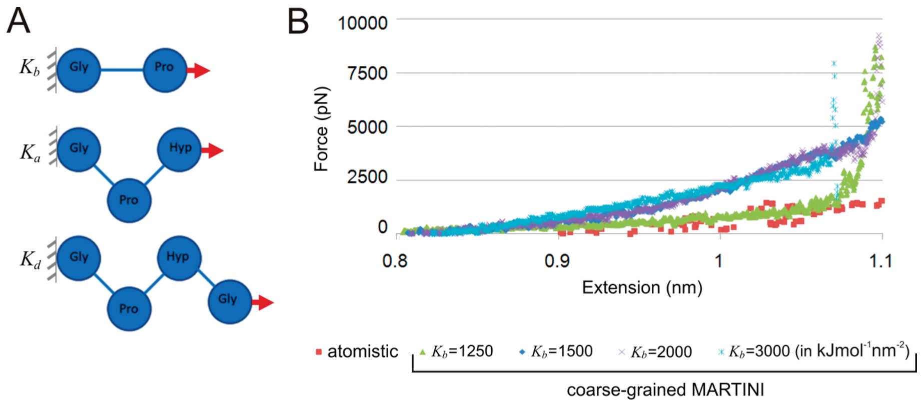


***Figure 2.*** Schematic of the setup used to calculate elastic constants for the coarse-grain force field (panel A) and force-extension plot of atomistic and coarse-grain peptides used to find the optimal value of bonding elastic constant (panel B), varying $K_b$ from 1250 kJmol$^{-1}$ nm$^{-2}$ up to 3000 kJ mol$^{-1}$ nm$^{-2}$. The bonding elastic constant ($K_b$) is obtained by pulling a glycine−proline peptide, the angle elastic constant ($K_a$) from a glycine−proline−hydroxiproline peptide, and the dihedral elastic constant ($K_d$) from a glycine−proline−hydroxiproline−glycine peptide. The molecules are considered both in coarse-grained and in atomistic forms and are subjected to steered molecular dynamics simulations. The values of the bond, angle, and dihedral elastic constant are then optimized in order to match the behavior of the atomistic model.

***Table 1.*** Backbone Bonded Parameters for the Collagen Triple Helix (Present Work) and for the Other Secondary Structures, as Included in the Original MARTINI Force Field[23]

| backbone | $d_b$ [nm] | $K_b$ [kJ mol$^{-1}$ nm$^{-2}$] | $\varphi_a$ [deg] | $K_a$ [kJ mol$^{-1}$] | $\Psi_d$ [deg] | $K_d$ [kJ mol$^{-1}$] |
|---|---|---|---|---|---|---|
| **collagen triple helix [present work]** | **0.365** | **1250** | **119.2** | **150** | **−89.3** | **100** |
| α-helix | 0.35 | 1250 | 96 | 700 | 60 | 400 |
| coil | 0.35 | 200 | 127 | 25 | | |
| extended | 0.35 | 1250 | 134 | 25 | 180 | 10 |
| turn | 0.35 | 500 | 100 | 25 | | |
| bend | 0.35 | 400 | 130 | 25 | | |

spring constants (that is, the parameters $K_b$, $K_a$, and $K_d$) are initially set equal to the values adopted in the original MARTINI force field for an α helix secondary structure[23] and then gradually changed in order to best fit the atomistic force-extension curves so that they show the same stiffnesses. The procedure is first performed on the dipeptide (glycine−proline) in order to find the optimal value of $K_b$. Once this parameter is set, we consider a glycine−proline−hydroxyproline peptide in order to find the value of $K_a$ that best approximates the behavior of the corresponding atomistic system. Finally, having fixed $K_b$ and $K_a$, the glycine−proline−hydroxyproline−glycine peptide is modeled in order to determine the value of $K_d$ (see Figure 2A). Figure 2B shows a plot of force versus deformation in the case of the two amino acid peptides, used to calculate the bonding elastic constant $K_b$ between two beads. The values that best replicate the atomistic behavior are $K_b$ = 1250 kJ mol$^{-1}$ nm$^{-2}$, $K_a$ = 150 kJ mol$^{-1}$, and $K_d$ = 100 kJ mol$^{-1}$. Table 1 shows an overview of all model parameters.

Due to coupling of bonded and nonbonded force field terms, the equilibrium geometrical features may in general differ from force field reference values. However, to address this issue, we monitor the equilibrium values and find no significant difference with respect to the reference values.

**3.3. Validation of the Extended MARTINI Model: Young's Modulus and Persistence Length of the Collagen Molecule.** In order to validate the extended MARTINI force field model, an 8-nm-long collagen-like molecule (similar to the PDB entry used to assess the geometrical features) is pulled under the same conditions previously described. The chosen molecule is [(glycine−proline−hydroxyproline)$_{10}$]$_3$, which represents an "ideal" reference collagen molecule that was used in several earlier molecular dynamics works.[14,17] The RMSD of the coarse grain peptide during equilibration is compared with the equivalent atomistic counterpart, showing that the models reach similar and stable configurations (see Figure 3A). Figure 3B shows a snapshot of both atomistic and coarse-

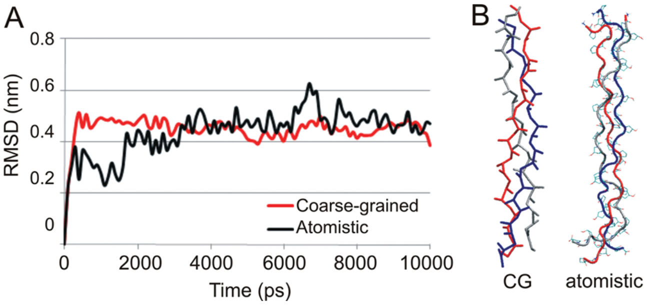


***Figure 3.*** Comparison of the atomistic and coarse-grained structures of a collagen molecule. Panel A displays the root mean squared displacement (RMSD) with respect to the starting structure during 10 ns of simulation time. Panel B shows snapshots of stable collagen triple helical structures in the coarse-grained (CG, left) and atomistic (right) representations. The images show that the triple helical structure is well-maintained in the CG representation.

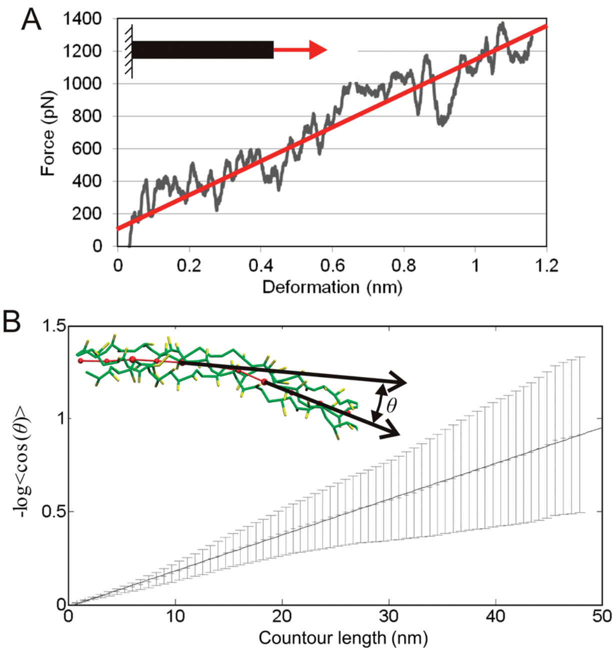


***Figure 4.*** Mechanical analysis of collagen molecules. Panel A: Force-deformation plot up to 15% strain (1.2 nm) for the pulling test of a [(glycine−proline−hydroxyproline)$_{10}$]$_3$ collagen-like peptide (straight line) and linear regression (dashed line) used to obtain the molecular elastic constant. Panel B: Plot of −log <cos(*θ*)> versus contour length calculated for the coarse-grain model of the first strand (≈ 60 nm) of the human collagen type I molecule. The linear regression of the data for the five different strands leads to an average persistence length of ~51.5 nm for human type I collagen. The inlay in panel B shows a schematic of the data analysis approach used to determine the collagen persistence length. The molecule is divided into segments (red lines) which are delimited by the center of mass (red balls) of the three equivalent glycine residues on the three collagen chains. The angle *θ*(*s*) is the angle between two segments that are separated by a contour length *s* (where *s* is the sum of the segments' lengths). The value of the angle *θ*(*s*) is obtained averaging over the 350 ns trajectory.

grained collagen peptides. The equilibrated coarse-grained collagen peptide is pulled using a deformation rate of 1 m/s. The Young's modulus of the coarse-grain collagen molecule is then calculated from the slope of the force-deformation plot, as described in previous works.[17,31] Force-extension plots similar to one performed on a coarse-grained molecule are shown in Figure 4A, together with a linear regression analysis used to obtain the stiffness of the collagen molecule (1052.78 ± 51.23 pN/nm) up to 15% strain. Assuming a cylindrical geometry, and considering a collagen molecular radius of 0.8 nm, the Young's modulus is obtained equal to 4.62 ± 0.41 GPa. Table 2 shows the values of Young's modulus obtained through different approaches, showing that our result agrees well with earlier findings.

Finally, the whole structure of the heterotrimeric type I collagen molecule is investigated in order to derive the persistence length. The full-length coarse-grained tropocollagen model is divided into five 60 nm peptide strands to reduce the computational cost. Then, the flexibility of the structures is analyzed from molecular dynamics trajectories of 350 ns each, obtaining the persistence length through the following expression:[35]

$$\log \langle \cos \theta \rangle = -\frac{s}{L_{\mathrm{p}}} \qquad (4)$$

As shown in the inlay in Figure 4B, the variable *θ* denotes the angle between two segments along the molecule separated by contour length *s* (that is, the sum of the segment lengths), and $L_{\mathrm{p}}$ is the persistence length, which provides information about a molecule's flexibility. In order to evaluate the expression defined in eq 4, the molecule is divided into several smaller segments. Each segment (as shown in the inlay in Figure 4B), is delimited by the center of mass of the three equivalent glycine residues belonging to the three different collagen chains. The plot of −log ⟨cos(*θ*)⟩ versus *s* is obtained, where the average of cos(*θ*) is calculated over the full molecular dynamics simulations trajectories. Figure 4B shows the plot of −log ⟨cos(*θ*)⟩ versus the contour length *s*. The value of the persistence length $L_{\mathrm{p}}$ is then obtained from the inverse of the slope after a linear regression of all data, giving a value of $L_{\mathrm{p}} = 51.5 \pm 0.6.7$ nm.

The persistence length found based on the coarse-grained model is close to the value obtained by Hofmann and co-workers,[36] who found a value of 57 ± 5 nm for collagen type I molecule through an electron microscopy analysis. On the other hand, a literature analysis shows that great variability in persistence length value can be found considering different kinds of experimental tests. The pioneering work in this analysis is represented by Utiyama and co-workers,[37] who considered sedimentation constants and the intrinsic viscosity of purified collagen molecules and calculated a value for the persistence length close to 130 nm. Saito and co-workers,[38] considering the hydrodynamic properties of collagen, derived the intrinsic viscosity and the sedimentation constant of collagen and, from these values, the persistence length, equal to 160−180 nm. In another experiment, Nestler and co-workers[39] measured the dynamic viscoelastic properties of dilute solutions of collagen molecules. From the obtained values of intrinsic viscosity and rotational relaxation time, they found a value for the persistence length of about 170 nm. In more recent work, Sun and co-workers[9] used optical tweezers in order to obtain, from force-extension plots fitted to the Marko−Siggia entropic elastic model,[40] the collagen persistence length, which was found to be 14.5 ±

***Table 2.*** Comparison of Young's Modulus of the Single Collagen Molecule Calculated Using Different Experimental and Computational Analysis

| type of analysis | Young's modulus (GPa) |
|---|---|
| X-ray diffraction[7] | ~3 |
| Brillouin light scattering[45] | ~5.1 |
| estimate based on persistence length[9] | 0.35−12 |
| estimate based on persistence length[36] | ~3 |
| single molecule stretching−atomistic modeling[13] | ~4.8 |
| single molecule stretching−reactive atomistic modeling[14] | ~7 |
| single molecule stretching−atomistic modeling[12] | ~2.4 |

***Table 3.*** Comparison of the Persistence Length ($L_p$) of Several Biopolymers

| polymer | configuration | $L_p$ |
|---|---|---|
| titin[46] | linear protein with two domains | ~15 nm |
| spectrin[47] | double-strand filament | 10−20 nm |
| collagen [optical tweezers][48] | triple helix | 14.5 ± 7.3 nm |
| collagen [MD bending simulations][16] | triple helix | 23.4 nm |
| **collagen [present work, coarse-grain model]** | **triple helix** | **51.5 ± 6.7 nm** |
| collagen [electron microscopy][3] | triple helix | 57 ± 5 nm |
| collagen [hydrodynamic properties][37,39,49] | triple helix | 130−180 nm |
| DNA[50] | double helix | ~50 nm |
| F-actin[35] | Filament | 0.5−15 $\mu$m |
| intermediate filaments[51] | 32 strands filament | ~1 $\mu$m |
| microtubules[52] | 13 protofilaments | 0.1−10 mm |

7.3 nm. Finally, earlier molecular simulation studies were used to perform bending tests on a collagen-like peptide, predicting the bending stiffness of the collagen molecule and, from this result, evaluate the persistence length of collagen to be in the range of 16−24 nm.[16] The variability that can be found in the literature for the collagen persistence length value depends on the used experimental setup, for example, electron micrographs or optical tweezers, on the experimental conditions (how the sample is extracted and prepared), on the method applied and the parameters used to fit experimental data with theoretical models. Despite the large variability of experimental values, the value obtained in the present work is in good agreement with the consideration of collagen as a semiflexible molecule with a rodlike configuration. A comparison with other biomolecules shows that the persistence length of biomolecules spans several order of magnitudes (see Table 3) and that, as for collagen, in some other cases, a very precise value cannot be established, but rather a range. In the case of microtubules, Pampaloni et al.[41] found that the persistence length value ranges from a few hundred micrometers up to a few millimeters. Soictin shows a certain variability depending on the kind of setup used to derive the persistence length, with values from 0.5 $\mu$m up to 15 $\mu$m.[42] Table 3 shows the values of the persistence length of collagen molecules obtained through different approaches, showing that our result agrees reasonably well with earlier findings, albeit tending to be larger than what has been computed on the basis of earlier full atomistic modeling.

The calculation of $L_p$ provides an alternative way to estimate the Young's modulus $Y$, since

$$Y = \frac{4L_p k_B T}{\pi r^4} \tag{5}$$

where $L_p$ is the persistence length, $k_B$ is Boltzmann's constant, $T$ is the absolute temperature, $r$ is the molecular radius, and $Y$ is the Young's modulus. Considering a temperature of 300 K, $r = 0.8$ nm, and applying eq 5, the Young's modulus $Y$ results to be 0.66 ± 0.08 GPa. The deviation between direct stretching and the estimation of Young's modulus via the persistence length data suggests that eq 5 may not hold, which may reflect the fact that a collagen molecule is not a structurally homogeneous molecule as assumed in the underlying continuum theory.

## 4. Discussion and Conclusion

The most important contribution of this article is the extension of the MARTINI coarse-grained force field to include parameters that allow the modeling of collagen molecules. The original MARTINI protein force field lacks parameters for hydroxyproline, which is a nonstandard amino acid but is found frequently in natural collagen molecules. Furthermore, the MARTINI force field can describe a variety of protein secondary structures but lacks some peculiar cases of secondary structures, such as the collagen triple helix, in its original form. These limitations are overcome in the extended MARTINI coarse-grained force field reported here, now enabling the application of coarse-grained MARTINI models to model collagenous protein materials. Due to the broad significance of collagenous tissues in biomechanics, biochemistry, and biology in general, the new model could find wide applicability in many future studies.

Several validation computations have been carried out to ensure that the extended MARTINI model can accurately describe key mechanical and biophysical parameters of collagen molecules. First, we considered the Young's modulus of a collagen molecule by simulating a short collagen-like peptide with the sequence [(glycine−proline−hydroxyproline)$_{10}$]$_3$. The collagen molecule was subjected to axial load by using steered molecular dynamics, and the study resulted in a Young's modulus value of 4.62 GPa, which is in good agreement with those obtained both through atomistic setups[12,13,15,43] and experimental analysis[44,7,8] (see Table 2). Second, we have computed the persistence length of collagen molecules, leading to $L_p = 51.5 \pm 0.6.7$ nm. We find reasonable agreement with results from earlier studies (see Table 3).

Due to its contour length of 300 nm, all-atom simulations with explicit solvent are prohibitive since they would require excessive computational resources due to the very large number of particles. The reduction by roughly a factor of 10 in the coarse-grained description provides a significant speedup that facilitates the direct simulation of much longer molecules. Furthermore, the typical time step used in classical molecular dynamics simulations (1−2 fs) allows the modeling only on the nanosecond time scale. However, the coarse-grained model enables one to use much longer time-scales on the order of 20−40 fs. Considering the combined effect of the reduction of the number of particles and the increased time step, the coarse-grained approach leads to a total speedup of 200−400 with respect to atomistic simulations.

With this significant computational speedup, the modeling framework reported here opens many possibilities for future studies, particularly at the scale of collagen fibrils and possibly fibers. We note that the MARTINI coarse grain approach is only valid when the phenomena under study do not involve changes of the secondary structure. In the context of collagen, only events that do not involve unfolding of the collagen triple helix can be correctly modeled using the MARTINI force field formulation presented here. While unfolding of molecules is likely to play an important role in large deformation and the fracture of collagenous tissues, the mechanical stresses experienced by collagen tissues under small mechanical loads may not lead to unfolding at the molecular level. This is because strain is distributed over several hierarchical levels (i.e., fibers, fibrils, molecules) and involves several concurrent mechanisms (fiber uncrimping, proteoglycan-mediated fibril sliding, molecular slippage, and molecular elongation).

The new method could be particularly useful to predict and analyze the structure of collagen fibrils and even collagen fibers. For example, the coarse-grain approach may facilitate the computational investigation of how changes in the sequence would influence the packing of collagen fibril, helping the understanding of the mechanisms underlying collagen-related diseases, such as *osteogenesis imperfecta*. This, together with the study of the interaction with other relevant biomolecules such as proteoglycans, will provide useful details for the understanding of structure−property relationships in the broader class of collagen materials and as such makes an important contribution to materiomics.

**Acknowledgment.** M.J.B. acknowledges support through a National Science Foundation CAREER award (CMMI-0642545). This research was supported in part through the MIT-Italy program (Progetto Rocca).

**Supporting Information Available:** A script to convert an atomistic PDB file to a coarse-grained PDB file (atom2cg.awk) as well as a script to generate MARTINI coarse-grained topologies (seq2itp.pl) are included. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Kadler, K. E.; Baldock, C.; Bella, J.; Boot-Handford, R. P. Collagens at a glance. *J. Cell Sci.* **2007**, *120*, 1955.

(2) Hyde, T. J.; Bryan, M. A.; Brodsky, B.; Baum, J. Sequence dependence of renucleation after a Gly mutation in model collagen peptides. *J. Biol. Chem.* **2006**, *281*, 36937–36943.

(3) Hoffmann, D.; Voss, T.; Kuhn, K.; Engel, J. Localization of flexible sites in thread-like molecules from electron micrographs. Comparison of interstitial, basement membrane and intima collagens. *J. Mol. Biol.* **1984**, 172.

(4) Cowin, S. C. The mechanical and stress adaptive properties of bone. *Ann. Biomed. Eng.* **1983**, *11*, 263–95.

(5) Srinivasan, M.; Uzel, S. G. M.; Gautieri, A.; Keten, S.; Buehler, M. J. Alport Syndrome mutations in type IV tropocollagen alter molecular structure and nanomechanical properties. *J. Struct. Biol.* **2009**, *168*, 503–510.

(6) Beck, K.; Chan, V. C.; Shenoy, N.; Kirkpatrick, A.; Ramshaw, J. A. M.; Brodsky, B. Destabilization of osteogenesis imperfecta collagen-like model peptides correlates with the identity of the residue replacing glycine. *P. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 4273–4278.

(7) Sasaki, N.; Odajima, S. Stress-strain curve and Young's modulus of a collagen molecule as determined by the X-ray diffraction technique. *J Biomech.* **1996**, *29*, 655–658.

(8) Bozec, L.; Horton, M. Topography and mechanical properties of single molecules of type I collagen using atomic force microscopy. *Biophys. J.* **2005**, *88*, 4223–4231.

(9) Sun, Y.; Luo, Z.; Fertala, A.; An, K. Direct quantification of the flexibility of type I collagen monomer. *Biochem. Biophys. Res. Commun.* **2002**, *295*, 382–386.

(10) Fratzl, P.; Weinkamer, R. Nature's hierarchical materials. *Prog. Mater. Sci.* **2007**, *52*, 1263–1334.

(11) Persikov, A. V.; Ramshaw, J. A.; Brodsky, B. Collagen model peptides: Sequence dependence of triple-helix stability. *Biopolymers* **2000**, *55*, 436–450.

(12) Vesentini, S.; Fitie, C. F. C.; Montevecchi, F. M.; Redaelli, A. Molecular assessment of the elastic properties of collagen-like homotrimer sequences. *Biomech. Model Mechanobiol.* **2005**, *3*, 224–234.

(13) Lorenzo, A. C.; Caffarena, E. R. Elastic properties, Young's modulus determination and structural stability of the tropocollagen molecule: a computational study by steered molecular dynamics. *J. Biomech.* **2005**, *38*, 1527–1533.

(14) Buehler, M. J. Atomistic and continuum modeling of mechanical properties of collagen: Elasticity, fracture and self-assembly. *J. Mater. Res.* **2006**, *21*, 1947–1961.

(15) Buehler, M. J. Nature designs tough collagen: Explaining the nanostructure of collagen fibrils. *P. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 12285–12290.

(16) Buehler, M. J.; Wong, S. Y. Entropic elasticity controls nanomechanics of single tropocollagen molecules. *Biophys. J.* **2007**, *93*, 37–43.

(17) Gautieri, A.; Vesentini, S.; Montevecchi, F. M.; Redaelli, A. Mechanical properties of physiological and pathological models of collagen peptides investigated via steered molecular dynamics simulations. *J. Biomech.* **2008**, *41*, 3073–3077, DOI:, 10.1016/j.jbiomech.2008.06.028.

(18) Ayton, G. S.; Noid, W. G.; Voth, G. A. Multiscale modeling of biomolecular systems: in serial and in parallel. *Curr. Opin. Struct. Biol.* **2007**, *17*, 192–198.

(19) Noid, W. G.; Chu, J. W.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C. The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.* **2008**, 128.

(20) Wang, Y.; Noid, W. G.; Liu, P.; Voth, G. A. Effective force coarse-graining. *Phys. Chem. Chem. Phys.* **2009**, *11*, 2002–15.

(21) Tozzini, V. Coarse-grained models for proteins. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144–150.

(22) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. The MARTINI force field: Coarse grained model for biomolecular simulations. *J. Phys. Chem. B* **2007**, *111*, 7812–7824.

(23) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S. J. The MARTINI coarse-grained force field: Extension to proteins. *J. Chem. Theory Comput.* **2008**, *4*, 819–834.

(24) Lopez, C. A.; Rzepiela, A. J.; de Vries, A. H.; Dijkhuizen, L.; Hunenberger, P. H.; Marrink, S. J. Martini Coarse-Grained Force Field: Extension to Carbohydrates. *J. Chem. Theory Comput.* **2009**, *5*, 3195–3210.

(25) Chetwynd, A. P.; Scott, K. A.; Mokrab, Y.; Sansom, M. S. CGDB: a database of membrane protein/lipid interactions by coarse-grained molecular dynamics simulations. *Mol. Membr. Biol.* **2008**, *25*, 662–9.

(26) Sansom, M. S.; Scott, K. A.; Bond, P. J. Coarse-grained simulation: a high-throughput computational approach to membrane proteins. *Biochem. Soc. Trans.* **2008**, *36*, 27–32.

(27) Treptow, W.; Marrink, S. J.; Tarek, M. Gating motions in voltage-gated potassium channels revealed by coarse-grained molecular dynamics simulations. *J. Phys. Chem. B* **2008**, *112*, 3277–82.

(28) Risselada, H. J.; Marrink, S. J. Curvature effects on lipid packing and dynamics in liposomes revealed by coarse grained molecular dynamics simulations. *Phys. Chem. Chem. Phys.* **2009**, *11*, 2056–67.

(29) Berendsen, H. J. C.; Vanderspoel, D.; Vandrunen, R. Gromacs - a message-passing parallel molecular-dynamics implementation. *Comput. Phys. Commun.* **1995**, *91*, 43–56.

(30) Vandrunen, R.; Vanderspoel, D.; Berendsen, H. J. C. Gromacs - a software package and a parallel computer for molecular-dynamics. *Abstr. Pap. Am. Chem. Soc.* **1995**, *209*, 49.

(31) Gautieri, A.; Buehler, M. J.; Redaelli, A. Deformation rate controls elasticity and unfolding pathway of single tropocollagen molecules. *J. Mech. Behav. Biomed. Mater.* **2009**, *2*, 130–137.

(32) Rainey, J.; Goh, M. An interactive triple-helical collagen builder. *Bioinformatics* **2004**, *20*, 2458–2459.

(33) Black, S. D.; Mould, D. R. Development of hydrophobicity parameters to analyze proteins which bear posttranslational or cotranslational modifications. *Anal. Biochem.* **1991**, *193*, 72–82.

(34) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(35) Chu, J. W.; Voth, G. A. Coarse-grained modeling of the actin filament derived from atomistic-scale simulations. *Biophys. J.* **2006**, *90*, 1572–1582.

(36) Hofmann, H.; Voss, T.; Kuhn, K.; Engel, J. Localization Of Flexible Sites In Thread-Like Molecules From Electron-Micrographs - Comparison Of Interstitial, Basement-Membrane And Intima Collagens. *J. Mol. Biol.* **1984**, *172*, 325–343.

(37) Utiyama, H.; Sakato, K.; Ikehara, K.; Setsuiye, T.; Kurata, M. Flexibility Of Tropocollagen From Sedimentation And Viscosity. *Biopolymers* **1973**, *12*, 53–64.

(38) Saito, T.; Iso, N.; Mizuno, H.; Onda, N.; Yamato, H.; Odashima, H. Semi-Flexibility Of Collagens In Solution. *Biopolymers.* **1982**, *21*, 715–728.

(39) Nestler, F. H. M.; Hvidt, S.; Ferry, J. D.; Veis, A. Flexibility Of Collagen Determined From Dilute-Solution Viscoelastic Measurements. *Biopolymers* **1983**, *22*, 1747–1758.

(40) Bustamante, C.; Marko, J. F.; Siggia, E. D.; Smith, S. Entropic Elasticity Of Lambda-Phage Dna. *Science.* **1994**, *265*, 1599–1600.

(41) Pampaloni, F.; Lattanzi, G.; Jonas, A.; Surrey, T.; Frey, E.; Florin, E. L. Thermal fluctuations of grafted microtubules provide evidence of a length-dependent persistence length. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 10248–10253.

(42) Mogilner, A.; Oster, G. Cell motility driven by actin polymerization. *Biophys. J.* **1996**, *71*, 3030–3045.

(43) Buehler, M. J. Atomistic modeling of elasticity, plasticity and fracture of protein crystals. *J. Comput. Theor. Nanosci.* **2006**, *3*, 670–683.

(44) Sasaki, N.; Odajima, S. Elongation mechanism of collagen fibrils and force-strain relations of tendon at each level of structural hierarchy. *J Biomech.* **1996**, *29*, 1131–1136.

(45) Cusack, S.; Miller, A. Determination Of The Elastic-Constants Of Collagen By Brillouin Light-Scattering. *J. Mol. Biol.* **1979**, *135*, 39–51.

(46) Higuchi, H.; Nakauchi, Y.; Maruyama, K.; Fujime, S. Characterization of beta-connectin (titin-2) from striated-muscle by dynamic light-scattering. *Biophys. J.* **1993**, *65*, 1906–1915.

(47) Svoboda, K.; Schmidt, C. F.; Branton, D.; Block, S. M. Conformation and elasticity of the isolated red-blood-cell membrane skeleton. *Biophys. J.* **1992**, *63*, 784–793.

(48) Sun, Y. L.; Luo, Z. P.; Fertala, A.; An, K. N. Direct quantification of the flexibility of type I collagen monomer. *Biochem. Biophys. Res. Commun.* **2002**, *295*, 382–386.

(49) Saito, T.; Iso, N.; Mizuno, H.; Onda, N.; Yamato, H.; Odashima, H. Semiflexibility of collagens in solution. *Biopolymers.* **1982**, *21*, 715–28.

(50) McCauley, M. J.; Williams, M. C. Mechanisms of DNA binding determined in optical tweezers experiments. *Biopolymers.* **2007**, *85*, 154–168.

(51) Mucke, N.; Kreplak, L.; Kirmse, R.; Wedig, T.; Herrmann, H.; Aebi, U.; Langowski, J. Assessing the flexibility of intermediate filaments by atomic force microscopy. *J. Mol. Biol.* **2004**, *335*, 1241–1250.

(52) Kikumoto, M.; Kurachi, M.; Tosa, V.; Tashiro, H. Flexural rigidity of individual Microtubules measured by a buckling force with optical traps. *Biophys. J.* **2006**, *90*, 1687–1696.

CT100015V

JCTC Journal of Chemical Theory and Computation

# Computational Study of Promising Organic Dyes for High-Performance Sensitized Solar Cells

David Casanova,*,† François P. Rotzinger,‡ and Michael Grätzel‡

*Institut de Química Teòrica i Computacional (IQTCUB), Universitat de Barcelona,
Martí i Franquès, 1-11, 08028 Barcelona, Spain, and Institut des Sciences et
Ingénierie Chimiques (ISIC), Laboratoire de Photonique et Interfaces, Ecole
Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland*

Received October 27, 2009

**Abstract:** The energy transition from the ground state to the first excited singlet of four organic dye candidates to be used as sensitizers in solar cells, D5, D7, D9, and D11, has been computationally explored and compared to experimental results with TDDFT (B3LYP, $\omega$B97, and $\omega$B97X functionals) and the CIS(D) and SOS-CIS(D) wave function based methods. The second-order perturbation correction to CI singles' excitation energies are superior to any TDDFT functional employed here. The performance of SOS-CIS(D) is especially interesting, being in good agreement with absorption spectra and having important computational savings. The best TDDFT results are obtained by the $\omega$B97X functional. Solvation effects on the excitation energies have been studied with three different models, i.e., the Onsager reaction field model, SS(V)PE, and SM8.

## Introduction

Dye-sensitized solar cells are alternatives to conventional semiconductor photovoltaic devices. With ruthenium polypyridyl dyes, up to 11% conversion efficiency at standard AM 1.5 sunlight has been achieved.[1,2] Metal-free organic dyes are promising alternatives because of their high extinction coefficients. With such compounds, efficiencies of ~4−8% have been reached so far.[3–6] To improve their yield, the light-absorption maximum needs to be shifted to the red, and the interaction of the excited (injecting) state of the dye with the conduction band of the semiconductor, on whose surface the dye is anchored, has to be improved.

One of the most recent studies in this direction corresponds to the analysis of four organic dye-sensitizers, e.g., 3-(5-(4-(diphenylamino)styryl)thiophen-2-yl)-2-cyanoacrylic acid (D5),[6] 3-(5-bis(4-(diphenylamino)styryl)thiophen-2-yl)-2-cyanoacrylic acid (D7), 5-(4-(bis(4-methoxyphenylamino)-styryl)thiophen-2-yl)-2-cyanoacrylic acid (D9), and 3-(5-bis(4,4′ -dimethoxydiphenylamino)styryl)thiophen-2-yl)-2-cyanoacrylic acid (D11). These molecules (Figure 1) were synthesized, anchored onto $TiO_2$, and tested in dye-sensitized solar cells, showing promising photon-to-current conversion efficiencies.[7]

From the experimental point of view, the syntheses of such sensitizers are quite demanding. Therefore, one would like to predict beforehand the potential properties of possible candidates in order to screen out molecules without the desired qualities and find those systems worth testing experimentally. To do that in a comprehensive and systematic manner, one could try to find some answers using the available quantum chemistry methods. Computational studies can be really helpful in providing some hints about the important aspects of the studied molecules, like the prediction of transition energies, oscillator strengths, or the electronic nature of the ground and excited states. But, first of all, these methods must be validated in the computation of the target molecules, and the comparison to experimental measurements is mandatory. These results should provide a standardized procedure to be applied in the study of similar sets of organic dyes.

In the computation of electronic excited states, a large variety of molecular quantum chemistry methods is available. The hierarchy of the approximations defines the properties of the models, classifying them at different levels of

* Corresponding author e-mail: davidcasanovacasas@ub.edu.

† Universitat de Barcelona.

‡ EPFL.

10.1021/ct100069q
Published on Web 03/15/2010

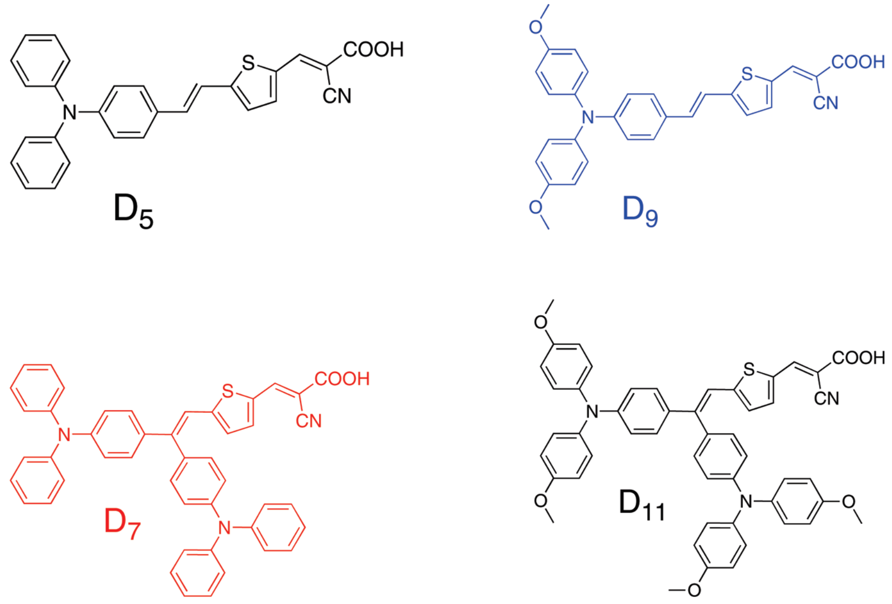


***Figure 1.*** Molecular structures of D5, D7, D9, and D11 dye sensitizers.

complexity. Of course, the chosen computation level will directly affect on the accuracy of the computed energy, but it also determines its computational cost. Although quite high accuracy is desired in the calculation of transition energies of organic dyes (ideally, one would like to achieve an accuracy better than $\sim$0.05 eV), computational demands are rather worrisome, especially due to the size of the molecules with the targeted properties. Thus, accurate models in the determination of single electron character excitation energies like the equation-of-motion coupled cluster (EOM-CC)[8–12] family of methods, the closely related linear response coupled-cluster (LR-CC),[13,14] or the symmetry-adapted cluster configuration interaction (SAC-CI)[15–17] methods cannot be considered in routine screenings of transition energies of such large molecules. These limitations become even stronger in multiconfigurational-based wave function methods, like multireference configuration interaction (MR-CI),[18] mutlireference coupled-cluster (MRCC),[19,20] or second-order perturbation correction to the complete active space self-consistent field (CASPT2[21] or MCQDPT2,[22,23] for example), which seem, in general, computationally prohibitive. Methods allowing for larger active spaces for the multiconfiguration SCF wave function (MCSCF) reference are available, in particular, second-order perturbation theory (PT2) based on restricted active space SCF (RASPT2)[24] or PT2 based on general MCSCF (GMCPT2).[25–27]

On the other hand, although time-dependent density functional theory (TDDFT)[28,29] represents a very attractive alternative, the standard density functionals exhibit a sizable delocalization and static correlation error,[30] as well as the self-interaction[31] error. The latter is responsible for large errors in long-range charge transfer transition energies[32–36] and becomes rather relevant in the kind of excitations occurring in molecular dyes. Therefore, one has to be very careful in drawing some conclusions derived from the application of such a methodology.

The present article reports on the computation of the lowest-energy electronic singlet-to-singlet transitions of metal-free organic dyes (Figure 1) used in dye-sensitized solar cells. These electronic excitations have been explored within the configuration interaction singles (CIS),[15] its second-order perturbation correction CIS(D)[37,38] and the scaled opposite spin version SOS-CIS(D),[39] and by TDDFT[28,29] using the popular B3LYP[40,41] functional and the recently developed long-range corrected (LC) $\omega$B97 and $\omega$B97X hybrid density functionals of Chai and Head-Gordon,[42] which substantially reduce standard long-range errors. The computed values are compared to experimental spectra. The available experimental UV−vis spectra were measured in ethanol solution. Since the charge transfer absorption bands can be very sensitive to the solvent polarity and hydrogen bonding, the effect of the solvent on excitation energies of D5, D7, D9, and D11 sensitizers was also taken into account.

The exposition of the present study is as follows. First, a detailed description of the computational tools employed is introduced. Second, a discussion of the obtained results based on the comparison to the available experimental data is presented. Finally, the main conclusions are exposed.

## Computational Details

Geometry optimizations of the D5, D7, D9, and D11 organic sensitizers (Figure 1) in the ground state and in ethanol solution have been performed with no symmetry restrictions at four different computational levels: B3LYP, $\omega$B97, and $\omega$B97X and at the scaled opposite spin MP2 (SOSMP2).[43,44] All geometry optimizations were computed in ethanol solution using the SM8 solvation model,[45] and with the 6-31G(d) basis set.[46,47] The discussed results correspond to B3LYP/6-31G(d) geometries unless indicated. The vertical excitation energies were calculated for all geometries by CIS,[15] CIS(D),[37,38] and SOS-CIS(D)[39] wave-function-based methods, and by the B3LYP, $\omega$B97, and $\omega$B97X functionals[42] within the TDDFT methodology. The 6-31+G(d) basis set has been used for all of the cases. The resolution-of-the-identity (RI) approximation[48,49] and the adoption of the

Laplace transform[50] in the integral evaluations were used in the CIS(D) and SOS-CIS(D) calculations. TDDFT energies were obtained within the Tamm−Dancoff (TDA) approximation.[51] All transitions were computed at the gas phase and ethanol solution. Three different solvation models were compared in the computation of transition energies: the Onsager reaction field model[52] up to the 15th multipole order expansion, Surface and Simulation of Volume Polarization for Electrostatics (SS(V)PE),[53–55] and the SM8 solvation model. In the Onsager and SS(V)PE cases, the solute−solvent interaction was described through the definition of a spherical cavity for the solute molecule. The radius of the cavities was defined as half the distance between the two outermost atoms of the solute molecule plus 1.4 Å to ensure that it was fully contained in the cavity. In the SM8 model, the solute cavities are defined by the superposition of nuclear-centered spheres whose sizes are determined by intrinsic atomic Coulomb radii. The employed value of the static dielectric constant for ethanol solvent was 24.85. Vertical excitation energies in solution have been computed within the equilibrium solvation approach. The 6-31G(d) basis was used in the comparison of the solvation models. Vibronic couplings have been neglected throughout the present study. All calculations were performed with the Q-Chem program.[56]

## Results and Discussion

The chemical structure of the studied dyes, i.e., D5, D7, D9, and D11, is composed of three modules: a triarylamine group acting as an electron donor, a thiophene which can be seen as an electron conductor, and an acceptor carrying the anchoring group, the cyano-acrylate fragment. This composition determines the general electronic properties of the D*n* molecules, and it further draws the boundaries of their spectroscopic behavior. In the D5 dye optimized geometry, the fragment constituted by the cyano-acrylate, the thiophene, and the chemically bonded benzene ring of the triarylamine group presents a planar space disposition, while the two benzene rings at one end of the molecule lie out of the plane (with dihedral angles to the plane between 60° and 90°). Besides the methoxy substitution in D9, there are almost no geometrical differences between D5 and D9 molecules, and only the dihedral angle between the two out of plane benzene rings of the triarylamine slightly increases from D5 to D9 (∼78° and ∼86°, respectively). The introduction of a second triarylamine group in D7 with respect to the D5 dye implies some relatively important structural modifications. The planarity observed in D5 and D9 cases is no longer preserved; i.e., any of the benzene rings appear coplanar to the thiophene and cyano-acrylate fragments. It is also worth mentioning the fact that, while one of the triarylamine groups has a rather linear disposition with respect to the thiophene and cyano-acrylate fragments, the second one appears in a quite perpendicular orientation with respect to the main molecular direction. As in the D5/D9 case, there are no relevant geometrical differences between D7 and D11 optimized geometries, and only a similar dihedral aperture between the two end benzene rings (in both triarylamine groups) is appreciated. In what follows, we will see how these different structural patterns induce slightly different spectroscopic properties between the studied sensitizers.

The most interesting excited electronic state of the D*n* dyes as molecular sensitizers is the lowest excited singlet. The ground to first excited state transition has a $\pi \rightarrow \pi^*$ nature with an important charge transfer character. Thus, although other allowed low-lying electronic transitions are present in the D*n* family,[7] we focus our efforts on the understanding of the mentioned lowest excited state.

**The One Electron Picture.** Before analyzing the set of results obtained with the several methodologies announced in the Computational Details section, we believe it is worth it to set a general frame that will help to understand the results. To that purpose and to qualitatively explain the role of different chemical substitutions, we analyze the properties of the molecular orbitals of the D*n* family. We are especially interested in understanding the different energy gaps between the frontier orbitals, which should give us some hints about the subtle differences in their spectroscopic behavior. In order to avoid redundancies, we restrict this discussion to the results obtained with the $\omega$B97X functional and the 6-31G(d) atomic basis set. Similar qualitative conclusions could be reached if we would pick the molecular orbitals obtained with another approach. We will also compare orbitals obtained from gas phase and ethanol solution within the SM8 model calculations.

Although the $\pi$-like HOMO of D5 is mainly localized at the triarylamine group, it noticeably expands toward the thiophene, probably due to the coplanarity of the fragments (Figure 2). On the other hand, the LUMO is basically localized at the thiopene and cyano-acrylate fragments, and only some small $\pi^*$ contribution on the closer benzene ring is obtained.

The p-electrons from the two methoxy oxygen atoms in D9 extend the $\pi$-system conjugation in the HOMO and HOMO−1, adding some extra antibonding interaction, which slightly destabilizes the two highest occupied orbitals. Meanwhile, the LUMO is basically unaffected by the methoxy substitution, resulting in an appreciable decrease of the energy gaps with respect to D5 (Figure 2).

The main part of the electronic density in the D7 (and D11) HOMO orbital is disposed along the triarylamine group, which is collinear to the thiophene and cyano-acrylate groups, and like in the D5 (D9) case, it delocalizes over the thiophene. On the other hand, the second triarylamine, almost perpendicular to the main molecular direction, has a less important contribution to the orbital. In addition, the small loss of planarity in D7 (D11) explained above seems to only very mildly affect the energies of the occupied orbitals, slightly pushing them to higher values. The strong localization of the LUMO orbital in D7 (and D11) avoids any major difference with respect to the D5 (D9) LUMO caused by the nonplanarity effect. This would explain the energy similarities between D5 and D7 (D9 and D11). Finally, the methoxy substitution in D7 to D11 conduces to a reduction of the HOMOs to LUMO energy separation, very similarly to the D5/D9 case. The present HOMOs and LUMOs resemble those of the similar DS-3, DS-4, and DS-5 dyes explored in a recent study.[57]

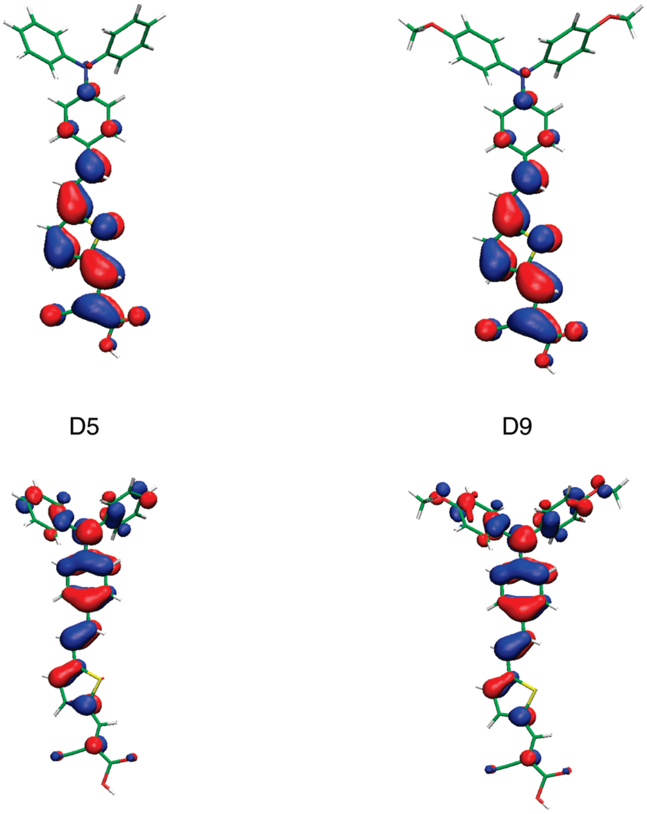


***Figure 2.*** Isodensity surface plots of HOMO (down) and LUMO (up) orbitals at the ωB97X/6-31G(d) computational level in the gas phase for the D5 (left) and D9 (right) molecular dyes. A 0.03 cutoff has been used in all cases.

When the molecular orbitals are computed within the SM8 model (ethanol as the solvent), very similar results are obtained. The most significant difference in the presence of the solvent is a ~0.1 eV stabilization of the LUMO due to the localized character of the orbital, which arises from hydrogen bonding and the polarity of the ethanol solvent ($\varepsilon = 24.85$).

**$S_0$ to $S_1$ Vertical Transition Energies.** Computed vertical energy gaps in the gas phase and ethanol are shown and compared to experimental absorption maxima in ethanol solution[7] in Table 1.

The solvent−solute interaction has been included at the structural level through geometry optimization of D*n* dyes within the SM8 model, and also at the electronic structure level, incorporating the environment effects in the computation of the ground and excited states. The computation of D*n* dyes in ethanol results in a moderate shift to lower absorption frequencies with respect to the gas phase calculations for all methods. This shift is almost constant for the CIS, ωB97, and ωB97X energies, about 0.05, 0.08, and 0.07 eV on average, respectively. The solvent effect in B3LYP is less constant but of similar magnitude (~0.10 eV). On the other hand, maybe due to the fact that the wave function has not been relaxed in the excited state in the presence of the solvent, the solvation redshift is considerably larger, about 0.15−0.21 eV, when CIS is corrected to the second-order perturbation theory in the CIS(D) and SOS-CIS(D) methods.

***Table 1.*** Transition Energies (in eV) of the D*n* Dyes in the Gas Phase and Ethanol Solution (SM8 solvation model) for All Studied Excited State Methods, Computed for the B3LYP/6-31G(d) SM8 Optimized Geometries with the 6-31+G(d) Basis Set

| method | D5 | D7 | D9 | D11 |
|---|---|---|---|---|
| | Vacuum | | | |
| B3LYP | 2.28 | 2.06 | 2.18 | 1.96 |
| ωB97 | 3.10 | 3.09 | 3.03 | 2.96 |
| ωB97X | 3.04 | 3.01 | 2.96 | 2.88 |
| CIS | 3.24 | 3.23 | 3.19 | 3.13 |
| SOS-CIS(D) | 2.75 | 2.72 | 2.64 | 2.53 |
| CIS(D) | 2.92 | 2.89 | 2.82 | 2.72 |
| | Ethanol, SM8 | | | |
| B3LYP | 2.16 | 2.07 | 1.98 | 1.88 |
| ωB97 | 3.04 | 3.01 | 2.95 | 2.88 |
| ωB97X | 2.97 | 2.94 | 2.87 | 2.81 |
| CIS | 3.20 | 3.18 | 3.13 | 3.07 |
| SOS-CIS(D) | 2.59 | 2.56 | 2.43 | 2.35 |
| SOS-CIS(D)[a] | 2.72 | 2.68 | 2.59 | 2.48 |
| CIS(D) | 2.77 | 2.75 | 2.62 | 2.54 |
| CIS(D)[a] | 2.89 | 2.84 | 2.77 | 2.66 |
| **experimental** | **2.81** | **2.81** | **2.68** | **2.71** |

[a] The solvent effects in CIS second-order corrected methods were obtained from the gas phase vs SM8 differences obtained in CIS.

Although the small redshift (about 0.1 eV) of the absorption band due to methoxy substitution between D5/D9 and D7/D11 is well recovered by B3LYP, as a result of the charge transfer nature of these π→π* transitions, vertical energy gaps to the first singlet are dramatically underestimated.[58-60] The loss of planarity in the larger D7 and D11 dyes decreases the delocalization of the orbitals, especially the HOMO, increasing the charge transfer character of the transition. This is reflected as a more severe underestimation of the transition by B3LYP (>0.7 eV) in the computation of D7 and D11 dyes. On the other hand, the LC functionals improve this ill behavior of standard functionals, and the computed energies are closer to experimental results. In this case, ωB97 and ωB97X overcorrect the long-range charge transfer transition failure of the standard functionals. The ωB97 functional systematically overestimates the lowest π→π* transitions in D*n* dyes by approximately 0.20 eV, while this effect is less pronounced with the inclusion of a small fraction of short-range Hartree−Fock exchange in ωB97X (the overestimation is reduced to 0.15 eV on average). As with B3LYP, both LC functionals are also able to correctly obtain the 0.1 eV redshift by methoxy substitution in D9 and D11.

CIS satisfactorily reproduces the 0.1 eV redshift by methoxy substitution, but as a result of the lack of dynamical correlation, all CIS transition energies present a systematic overestimation with a mean absolute error (MAE) of 0.40 eV with respect to experimental absorption maxima, which is a common feature of the CIS method in the computation of valence excitations of organic molecules. The inclusion of dynamical correlation through second-order perturbation theory noticeably corrects the overestimation in CIS, decreasing the transition energies in CIS(D) and SOS-CIS(D) by 0.48 and 0.67 eV on average, respectively. This correction drives to CIS(D) energies which are very close to experimental absorption maxima (MAE = 0.08 eV) but is a bit large in SOS-CIS(D). These results seem to indicate that, although it was previously shown[39] that only considering the scaled opposite spin component of the second order

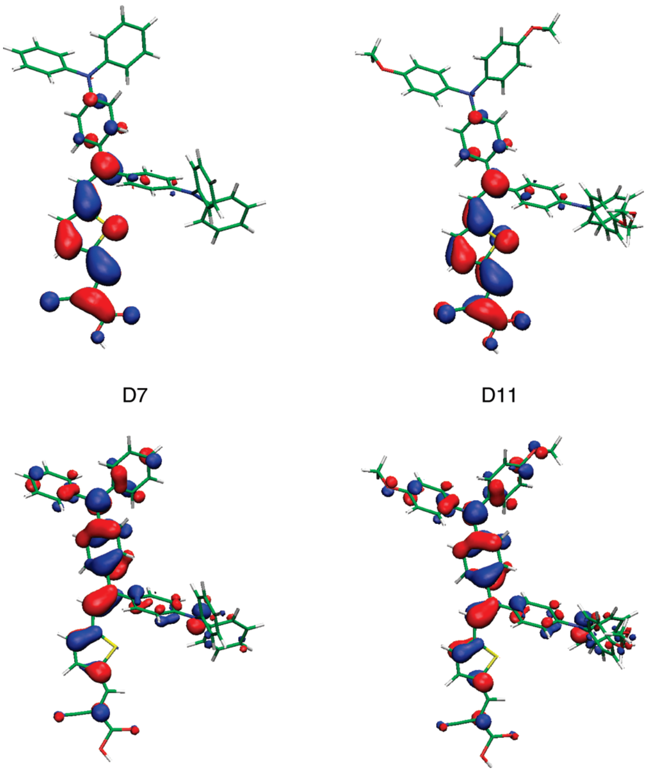


***Figure 3.*** Isodensity surface plots of HOMO (down) and LUMO (up) orbitals at the $\omega$B97X/6-31G(d) computational level in the gas phase for the D7 (left) and D11 (right) molecular dyes. A 0.03 cutoff has been used in all cases

***Table 2.*** Basis Set Dependence of the Transition Energies (in eV) of the D*n* Dyes in Gas Phase for All Studied Excited State Methods[a]

| method | 6-31G | 6-31G(d) | 6-31+G | 6-31+G(d) |
|---|---|---|---|---|
| | | D5 | | |
| $\omega$B97 | 3.24 | 3.18 | 3.16 | 3.10 |
| $\omega$B97X | 3.16 | 3.11 | 3.08 | 3.04 |
| CIS | 3.35 | 3.32 | 3.28 | 3.24 |
| SOS-CIS(D) | 3.05 | 2.85 | 2.95 | 2.75 |
| CIS(D) | 3.13 | 3.02 | 3.02 | 2.92 |
| | | D7 | | |
| $\omega$B97 | 3.21 | 3.16 | 3.14 | 3.09 |
| $\omega$B97X | 3.13 | 3.08 | 3.06 | 3.01 |
| CIS | 3.33 | 3.29 | 3.28 | 3.23 |
| SOS-CIS(D) | 3.02 | 2.81 | 2.92 | 2.72 |
| CIS(D) | 3.09 | 2.98 | 3.00 | 2.89 |
| | | D9 | | |
| $\omega$B97 | 3.16 | 3.11 | 3.09 | 3.03 |
| $\omega$B97X | 3.08 | 3.04 | 3.01 | 2.96 |
| CIS | 3.31 | 3.27 | 3.23 | 3.19 |
| SOS-CIS(D) | 2.94 | 2.74 | 2.84 | 2.64 |
| CIS(D) | 3.03 | 2.92 | 2.92 | 2.82 |
| | | D11 | | |
| $\omega$B97 | 3.07 | 3.02 | 3.01 | 2.96 |
| $\omega$B97X | 2.99 | 2.95 | 2.93 | 2.88 |
| CIS | 3.23 | 3.19 | 3.17 | 3.13 |
| SOS-CIS(D) | 2.82 | 2.62 | 2.72 | 2.53 |
| CIS(D) | 2.91 | 2.81 | 2.81 | 2.72 |

[a] All energies were computed on the B3LYP/6-31G(d) optimized geometries in ethanol (SM8).

correction of CIS slightly increases the perturbation correction and improves the accuracy of the computed valence $\pi \rightarrow \pi^*$ transitions in the gas phase with respect to CIS(D), the large solvatochromic redshifts (~0.2 eV) in the perturbationally corrected methods compensate for the slight to high excitation energies in CIS(D) but result in a systematic underestimation in SOS-CIS(D) (MAE = 0.27 eV). When the solvation correction coming from CIS is added to the SOS-CIS(D) in gas phase (Table 1), excitation energies are obtained in substantially better agreement with the experimental values, with a MAE = 0.14 eV underestimation. The same treatment of the solvation effects slightly improves CIS(D) energies (MAE = 0.06) in CIS(D). Both second-order corrected methods keep, in general, the same transition energy shift to lower frequencies between D5/D9 and D7/D11 obtained by CIS.

**$S_0$ to $S_1$ Transition Amplitudes.** The transition amplitudes obtained from the TDDFT and CIS methods give a quantitative description of the orbital contributions in the excited state wave function. In the D*n* family, the energetically lowest absorptions stem from $\pi \rightarrow \pi^*$ transitions and can be mainly described as HOMO to LUMO electronic promotions (see isodensity surface plots in Figures 2 and 3). For all explored methods but B3LYP, this amplitude, in the gas phase, constitutes about 60−70% of the final state, whereas HOMO−1 to LUMO in D5 and D9 and HOMO−2 to LUMO in D7 and D11 are responsible for 20−25%. Other possible contributions like HOMO to LUMO+1 are much less important. These compositions do not suffer any significant modification when obtained in ethanol, either using the Onsager, SS(V)PE, or SM8 solvation models analyzed below.

On the other hand, the HOMO to LUMO contribution in B3LYP accounts for ~95% of the transition of D5 and D9 dyes, whereas in D7 and D11, the electronic transit is built up from similar weights of the HOMO−1 and HOMO to LUMO excitations. When the ethanol effects are introduced, D5 and D9 transition amplitudes remain very similar to the gas phase, while becoming ~70% and ~30% for HOMO−1 and HOMO to LUMO for D7 and D11 molecules. This different behavior in B3LYP is symptomatic of the failures already present in the important underestimation of excitation energies.

**Transition Energy Dependence on the Basis Set.** The election of a basis set in the computation of molecular electronically excited states becomes crucial if one intends to obtain accurate results.[61−63] In this section, we explore how excitation energies to the lowest singlet in D*n* dyes vary with the employed basis. We do not pretend to produce a benchmark study, but we rather want to provide some hints of the importance of the basis set in computing the $S_0$ to $S_1$ transitions. Because of the failure of B3LYP to quantitatively capture the lowest singlet excited state of D*n* dyes, we have not included it in this study. The interested reader in the B3LYP excitation energies dependence on the basis set is referred to the Supporting Information (Tables S1 and S2).

As a result of the lack of polarization and diffusion functions, the excitations computed with the 6-31G basis (Table 2) are substantially higher compared to those with the larger 6-31+G(d) basis. This difference is on the order of 0.1 eV for CIS and TDDFT and even larger for CIS(D) and SOS-CIS(D), about 0.2 and 0.3 eV, respectively. When

***Table 3.*** Molecular Geometry Dependence of the Transition Energies (in eV) of the D*n* Dyes in Ethanol Solution (SM8 model) for All Studied Excited State Methods[a]

| method | optimized geometry | | | |
|---|---|---|---|---|
| | B3LYP | ωB97 | ωB97X | SOSMP2 |
| D5 ($\Delta E_{\mathrm{expt}} = 2.81$ eV) | | | | |
| ωB97 | 3.04 | 3.39 | 3.35 | 3.36 |
| ωB97X | 2.97 | 3.29 | 3.26 | 3.27 |
| CIS | 3.20 | 3.57 | 3.53 | 3.53 |
| SOS-CIS(D) | 2.59 | 3.09 | 3.03 | 3.03 |
| SOS-CIS(D)[b] | 2.72 | 3.15 | 3.11 | 3.10 |
| CIS(D) | 2.77 | 3.19 | 3.15 | 3.15 |
| CIS(D)[b] | 2.89 | 3.25 | 3.21 | 3.21 |
| D7 ($\Delta E_{\mathrm{expt}} = 2.81$ eV) | | | | |
| ωB97 | 3.01 | 3.39 | 3.35 | 3.32 |
| ωB97X | 2.94 | 3.31 | 3.28 | 3.24 |
| CIS | 3.18 | 3.60 | 3.56 | 3.52 |
| SOS-CIS(D) | 2.56 | 3.06 | 3.02 | 2.98 |
| SOS-CIS(D)[b] | 2.68 | 3.16 | 3.11 | 3.05 |
| CIS(D) | 2.75 | 3.18 | 3.15 | 3.11 |
| CIS(D)[b] | 2.84 | 3.25 | 3.22 | 3.17 |
| D9 ($\Delta E_{\mathrm{expt}} = 2.68$ eV) | | | | |
| ωB97 | 2.95 | 3.32 | 3.27 | 3.27 |
| ωB97X | 2.87 | 3.22 | 3.18 | 3.18 |
| CIS | 3.13 | 3.53 | 3.48 | 3.47 |
| SOS-CIS(D) | 2.43 | 2.99 | 2.93 | 2.91 |
| SOS-CIS(D)[b] | 2.59 | 3.07 | 3.01 | 3.00 |
| CIS(D) | 2.62 | 3.09 | 3.04 | 3.03 |
| CIS(D)[b] | 2.77 | 3.16 | 3.12 | 3.10 |
| D11 ($\Delta E_{\mathrm{expt}} = 2.71$ eV) | | | | |
| ωB97 | 2.88 | 3.28 | 3.24 | 3.26 |
| ωB97X | 2.81 | 3.19 | 3.15 | 3.18 |
| CIS | 3.07 | 3.49 | 3.46 | 3.47 |
| SOS-CIS(D) | 2.35 | 2.90 | 2.85 | 2.88 |
| SOS-CIS(D)[b] | 2.48 | 2.98 | 2.94 | 2.94 |
| CIS(D) | 2.54 | 3.01 | 2.97 | 3.01 |
| CIS(D)[b] | 2.66 | 3.08 | 3.04 | 3.06 |

[a] All energies were computed with the 6-31+G(d) basis on the (B3LYP, ωB97, ωB97X, SOSMP2)/6-31G(d) optimized geometries in ethanol (SM8). [b] The solvent effects in CIS second-order corrected methods were obtained from the gas phase vs SM8 differences obtained in CIS.

polarization functions are included, the excitation energies are systematically shrunk. This lowering is about 0.05 eV for all transitions computed by CIS and TDDFT, whereas this effect is twice as large in CIS(D) and four times larger (0.20 eV) in SOS-CIS(D). The reduction of transition energies by diffuse functions, i.e., the 6-31+G column in Table 2, has a similar magnitude to the one obtained in 6-31G(d), except in SOS-CIS(D), where, as in CIS(D), the effect accounts for a ~0.1 eV decrease in the excitation energies. These comparisons from Table 2 indicate that both polarization and diffuse functions substantially contribute to the final states, diminishing the $S_0$ to $S_1$ gaps by 0.10 to 0.30 eV.

Very similar results are obtained when excitation energies are computed for the different basis sets and in the presence of ethanol as a solvent (see Table S2 of Supporting Information).

**Transition Energy Dependence on the Geometry Optimization.** The molecular geometry used in electronic structure calculations can be a key factor when computing energies for ground and excited states. In Table 3, we compare the D*n* vertical transition energies in ethanol obtained from four different geometry optimization levels, i.e., B3LYP, ωB97, ωB97X, and SOSMP2.

The best performance of computed vertical energies is obtained when B3LYP is employed in the optimization of the ground state. Meanwhile, ωB97, ωB97X, and SOSMP2 geometries produce too large excitation energies in all cases. Compared to B3LYP-geometry results, transitions from ωB97X geometries are 0.30 to 0.50 eV higher. This difference is slightly increased (~0.05 eV) when ωB97 geometries are employed. Excitations from SOSMP2 optimized geometries are very similar to the ones obtained from ωB97X. These results have to be (only) seen as a consequence of a more favorable error compensation in B3LYP/6-31G(d) optimized geometries in the computation of D*n* $S_0$ to $S_1$ transitions.

Gas phase and ethanol solution (including B3LYP transitions) energy comparisons are presented as Supporting Information (Tables S3 and S4).

**Transition Energy Dependence on the Solvation Model.** Here, we perform a systematic analysis over three possible approaches to account for the solute−solvent interaction, i.e., the Onsager reaction field model[52] up to the 15th multipole order expansion, Surface and Simulation of Volume Polarization for Electrostatics (SS(V)PE),[53–55] and the SM8 solvation model,[45] in the computation of the D*n* dyes' excitation energies in ethanol solution. The larger energy stabilization of the LUMOs when the D*n* electronic structure is computed in solution makes us expect smaller energy gaps when the ethanol presence is modeled. The comparison of D*n* transition energies computed in the gas phase and with the three different solvation models is shown in Table 4.

In the Onsager model, specific electrostatic solute−solvent interactions are not treated, and the solute is enclosed in a spherical cavity. Most likely due to the neglect of specific electrostatic interactions, the excitation energies are very close to the gas phase results. In general, there is a small redshift (≤0.07 eV) with respect to gas phase energies for D5, D7, and D11. On the other hand, the D9 transition energies suffer similar magnitude frequency displacement but to larger energies. This behavior is common to all wave function-based and TDDFT explored methods. The SS(V)PE model treats electrostatic solute−solvent interactions by solving Poisson's equation at the cavity surface defined as the isodensity surface or a spherical cavity (dispersion and cavitation are neglected). Since the computations could not be converged for isodensity surfaces, solvation had to be computed using a spherical cavity. In this case, the electrostatic interactions are not localized properly. Surprisingly, SS(V)PE transition energies for the smaller dyes (D5 and D9) are slightly higher, between 0.03 and 0.28 eV, than the gas phase computed energies. Meanwhile, there is a small decrease in the D7 and D11 molecules, although the difference from the gas phase values is no larger than 0.07 eV. Since hydrogen bonding (with the ethanol solvent) is mainly electrostatic, the neglect of electrostatic interactions (Onsager) or its approximate treatment (SS(V)PE with a spherical cavity) is responsible for the errors. In contrast,

***Table 4.*** Transition Energies (in eV) of the D*n* Dyes in Ethanol Solution for All Studied Excited State Methods[a]

| method | vacuum | Onsager | SS(V)PE | SM8 |
|---|---|---|---|---|
| D5 ($\Delta E_{expt} = 2.81$ eV) | | | | |
| B3LYP | 2.34 | 2.31 | 2.42 | 2.20 |
| ωB97 | 3.18 | 3.19 | 3.25 | 3.12 |
| ωB97X | 3.11 | 3.12 | 3.18 | 3.04 |
| CIS | 3.32 | 3.31 | 3.35 | 3.25 |
| SOS-CIS(D) | 2.85 | 2.84 | 2.96 | 2.59 |
| SOS-CIS(D)[b] | | 2.84 | 2.88 | 2.78 |
| CIS(D) | 3.02 | 3.01 | 3.13 | 2.79 |
| CIS(D)[b] | | 3.02 | 3.05 | 2.96 |
| D7 ($\Delta E_{expt} = 2.81$ eV) | | | | |
| B3LYP | 2.11 | 2.05 | 2.06 | 1.97 |
| ωB97 | 3.16 | 3.13 | 3.13 | 3.09 |
| ωB97X | 3.08 | 3.05 | 3.05 | 3.01 |
| CIS | 3.29 | 3.27 | 3.27 | 3.23 |
| SOS-CIS(D) | 2.81 | 2.74 | 3.02 | 2.57 |
| SOS-CIS(D)[b] | | 2.79 | 2.79 | 2.75 |
| CIS(D) | 2.98 | 2.91 | 2.91 | 2.76 |
| CIS(D)[b] | | 2.96 | 2.96 | 2.92 |
| D9 ($\Delta E_{expt} = 2.68$ eV) | | | | |
| B3LYP | 2.23 | 2.27 | 2.46 | 2.08 |
| ωB97 | 3.11 | 3.16 | 3.26 | 3.05 |
| ωB97X | 3.04 | 3.09 | 3.19 | 2.97 |
| CIS | 3.27 | 3.29 | 3.35 | 3.20 |
| SOS-CIS(D) | 2.74 | 2.81 | 3.02 | 2.48 |
| SOS-CIS(D)[b] | | 2.76 | 2.82 | 2.67 |
| CIS(D) | 2.92 | 2.99 | 3.18 | 2.69 |
| CIS(D)[b] | | 2.94 | 3.00 | 2.86 |
| D11 ($\Delta E_{expt} = 2.71$ eV) | | | | |
| B3LYP | 2.01 | 1.96 | 1.97 | 1.83 |
| ωB97 | 3.02 | 2.99 | 2.99 | 2.95 |
| ωB97X | 2.95 | 2.92 | 2.91 | 2.87 |
| CIS | 3.19 | 3.17 | 3.16 | 3.12 |
| SOS-CIS(D) | 2.62 | 2.56 | 2.55 | 2.38 |
| SOS-CIS(D)[b] | | 2.60 | 2.64 | 2.56 |
| CIS(D) | 2.81 | 2.75 | 2.74 | 2.58 |
| CIS(D)[b] | | 2.79 | 2.78 | 2.74 |

[a] Solvation model dependence computed for the B3LYP/6-31G(d) SM8 optimized geometries with the 6-31G(d) basis set. [b] The solvent effects in CIS second order corrected methods were obtained from the gas phase vs SS(V)PE and SM8 differences obtained in CIS, respectively.

electrostatic interactions are treated with SM8, which is based on the generalized Born method for electrostatics augmented with atomic surface tensions for first-solvation-shell effects, and the deviation of the electrostatics from what can be calculated using only the bulk dielectric constant. Furthermore, cavitation and dispersion are included. Thus, all SM8 computed transition energies show a moderate redshift with respect to gas phase values. This behavior is almost constant (between 0.06 and 0.08 eV) for all transitions computed by TDDFT with the ωB97 and ωB97X functionals and the CIS method, and a little bit larger in B3LYP (~0.15 eV). As has been discussed above, the solvation redshift is considerably larger, about 0.25 eV, when CIS is corrected to the second order in the CIS(D) and SOS-CIS(D) methods. When the solvation correction coming from CIS is added to the SOS-CIS(D) in the gas phase, excitation energies are obtained in much better agreement with the experimental values.

The more accurate results obtained with SM8 are due to the approximate treatment of hydrogen bonding with the solvent. The present organic dyes are far from being spherically shaped, but the Onsager and SS(V)PE solvation models employed here require the dyes to be surrounded by a spherical cavity. As shown, computations with shape-adapted cavities, as it is done in SM8, seem much more convenient in these cases.

## Conclusions

The ground and first excited singlet states of the D5, D7, D9, and D11 molecular dyes have been studied. Computed excitation energies between them within TDDFT and wave-function-based methods have been compared to experimental absorption maxima, and the roles of different geometrical features and chemical substitutions have been discussed.

The smaller D5 and D9 dyes have a rather linear geometry and present localized HOMOs and LUMOs, especially the latter, at the two different molecular ends. The electron−hole separation in the electronic transition to the excited singlet results in a charge transfer nature of the excitation. The substitution of a second triarylamine group in D7 and D11 does not change the general picture of the electronic transition. The methoxy substitution in D9 and D11 produces a decrease of the excitation due to an extra antibonding interaction in the HOMO caused by the p-orbital of oxygen atoms with the benzene ring.

As far as the agreement with experimental results of transition energies to the first excited singlet is concerned, the CIS(D) method is superior to the LC ωB97 and ωB97X functionals. The scaled opposite spin version of CIS(D) represents an attractive alternative to CIS(D), especially due to its lower computational cost, although one must be careful when the solute−solvent interactions are included. Within the TDDFT realm, the ωB97X approach is the most promising tested functional in the computation of this kind of molecular dyes. On the other hand, B3LYP should not be the chosen method for routine screening of excitation energies of such molecules. The charge transfer nature of the first excited singlet produces catastrophic results when standard functionals, such as B3LYP, are used, which could drive us to wrong conclusions. Moreover, the magnitude of this underestimation, as has been shown between D5/D9 and D7/D11, strongly depends on the molecular characteristics, making this functional inappropriate for general comparisons.

The variation in the computed energy gaps due to different basis sets has been studied for all wave function-based and TDDFT employed methods. Similar contributions were found from polarization and diffuse functions when comparing 6-31G, 6-31+G, 6-31G(d), and 6-31+G(d) results. Transition energy dependence on the molecular geometry has been explored using four different optimization approaches. The best results were obtained when ground state geometries were optimized by the B3LYP functional. The use of ωB97, ωB97X, or SOSMP2 geometries drives to overestimation of the energy gaps. Finally, the role of ethanol as a solvent has been taken into account in the computation of optimized geometries and electronic structure calculations for three different solvation models. SM8 reproduces the expected solvent-induced redshift in the vertical excitation energies correctly. The deficiencies in the two other explored solvation models, i.e., Onsager and SS(V)PE, are due to the neglect or too approximate treatment of specific electrostatic solute–solvent interactions.

**Acknowledgment.** Financial support for this work was provided by the Spanish Ministerio de Ciencia e Innovación through project CTQ2008-06670-C02-02/BQU, the Generalitat de Catalunya through project 2009SGR-1459, and the Office of Basic Energy Sciences of the U.S. Department of Energy through the LBL Ultrafast Center, with additional support from the Director, Office of Energy Research, Office of Basic Energy Sciences, Chemical Sciences Division of the U.S. Department of Energy under Contract DE-AC0376SF00098, and supercomputer time from NERSC. D.C. gratefully acknowledges the Ramón y Cajal program for financial support and Professor Martin Head-Gordon for his comments, suggestions, and fruitful discussions on several of the topics presented in this work.

**Supporting Information Available:** Tables giving basis set and molecular geometry dependences. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Nazeeruddin, M. K.; De Angelis, F.; Fantacci, S.; Selloni, A.; Viscardi, G.; Liska, P.; Ito, S.; Takeru, B.; Grätzel, M. *J. Am. Chem. Soc.* **2005**, *127*, 16835.

(2) Grätzel, M. *J. Photochem. Photobiol., A* **2004**, *164*, 3.

(3) Hara, K.; Sato, T.; Katoh, R.; Furube, A.; Ohga, Y.; Shinpo, A.; Suga, S.; Sayama, K.; Sugihara, H.; Arakawa, H. *J. Phys. Chem. B* **2003**, *107*, 597.

(4) Horiuchi, T.; Miura, H.; Sumioka, K.; Uchida, S. *J. Am. Chem. Soc.* **2004**, *126*, 12218.

(5) Kim, S.; Lee, J. K.; Kang, S. O.; Ko, J.; Yum, J. H.; Fantacci, S.; De Angelis, F.; Di Censo, D.; Nazeeruddin, M. K.; Grätzel, M. *J. Am. Chem. Soc.* **2006**, *128*, 16701.

(6) Hagberg, D. P.; Edvinsson, T.; Marinado, T.; Boschloo, G.; Hagfeldt, A.; Sun, L. *Chem. Commun.* **2006**, 2245.

(7) Hagberg, D. P.; Yum, J.-H.; Lee, H.; De Angelis, F.; Marinado, T.; Karlsson, K. M.; Humphry-Baker, R.; Sun, L.; Hagfeldt, A.; Grätzel, M.; Nazeeruddin, M. K. *J. Am. Chem. Soc.* **2008**, *130*, 6259.

(8) Rowe, D. J. *Rev. Mod. Phys.* **1968**, *40*, 153.

(9) Stanton, J. F.; Bartlett, R. J. *J. Chem. Phys.* **1993**, *98*, 7029.

(10) Emrich, K. *Nucl. Phys. A* **1981**, *351*, 379.

(11) Geertsen, J.; Rittby, M.; Bartlett, R. J. *Chem. Phys. Lett.* **1989**, *164*, 57.

(12) Levchenko, S. V.; Krylov, A. I. *J. Chem. Phys.* **2004**, *120*, 175.

(13) Monkhorst, H. J. *Int. J. Quantum Chem., Symp.* **1977**, *11*, 421.

(14) Sekino, H.; Bartlett, R. J. *Int. J. Quantum Chem.* **1984**, *26*, 255.

(15) Janet, E. D. B.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1971**, *55*, 2236.

(16) Nakatsuji, H.; Hirao, K. *J. Chem. Phys.* **1978**, *68*, 2053.

(17) Nakatsuji, H. *Chem. Phys. Lett.* **1991**, *177*, 331.

(18) Siegbahn, P. E. M. In *Methods in Computational Molecular Physics*; Diercksen, G. H. F., Ed.; Reidel: Dordrecht, The Netherlands, 1983; p 189.

(19) *Recent Advances in Multi-reference Methods*; World Scientific: Singapore, 1999.

(20) Mukherjee, D.; Pal, S.; Per-Olov, L. In *Advanced Quantum Chemistry*;Academic Press: New York, 1989; Vol 20, p 291.

(21) Roos, B. O.; Andersson, K.; Fülscher, M. P. *Chem. Phys. Lett.* **1992**, *192*, 5.

(22) Nakano, H. *J. Chem. Phys.* **1993**, *99*, 7983.

(23) Nakano, H. *Chem. Phys. Lett.* **1993**, *207*, 372.

(24) Malmqvist, P. Å.; Pierloot, K.; Shahi, A. R. M.; Cramer, C. J.; Gagliardi, L. *J. Chem. Phys.* **2008**, *128*, 204109.

(25) Nakano, H.; Uchiyama, R.; Hirao, K. *J. Comput. Chem.* **2002**, *23*, 1166.

(26) Ebisuzaki, R.; Watanabe, Y.; Nakano, H. *Chem. Phys. Lett.* **2007**, *442*, 164.

(27) Miyajima, M.; Watanabe, Y.; Nakano, H. *J. Chem. Phys.* **2006**, *124*, 044101.

(28) Gross, E. K. U.; Kohn, W. *Phys. Rev. Lett.* **1985**, *55*, 2850.

(29) Runge, E.; Gross, E. K. U. *Phys. Rev. Lett.* **1984**, *52*, 997.

(30) Cohen, A. J.; Mori-Sanchez, P.; Yang, W. *Science* **2008**, *321*, 792.

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

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

(33) Sobolewski, A. L.; Domcke, W. *Chem. Phys.* **2003**, *294*, 73.

(34) Dreuw, A.; Weisman, J. L. *J. Chem. Phys.* **2003**, *119*, 2943.

(35) Dreuw, A.; Head-Gordon, M. *J. Am. Chem. Soc.* **2004**, *126*, 4007.

(36) Dreuw, A.; Fleming, G. R.; Head-Gordon, M. *Phys. Chem. Chem. Phys.* **2003**, *5*, 3247.

(37) Head-Gordon, M.; Rico, R. J.; Oumi, M.; Lee, T. J. *Chem. Phys. Lett.* **1994**, *219*, 21.

(38) Head-Gordon, M.; Grana, A. M.; Maurice, D.; White, C. A. *J. Phys. Chem.* **1995**, *99*, 14261.

(39) Rhee, Y. M.; Head-Gordon, M. *J. Phys. Chem. A* **2007**, *111*, 5314.

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

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

(42) Chai, J.-D.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *128*, 084106.

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

(44) Lochan, R. C.; Shao, Y.; Head-Gordon, M. *J. Chem. Theory Comput.* **2007**, *3*, 988.

(45) Marenich, A. V.; Olson, R. M.; Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 2011.

(46) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1972**, *56*, 2257.

(47) Hariharan, P. C.; Pople, J. A. *Theor. Chem. Acc.* **1973**, *28*, 213.

(48) Christof, H.; Florian, W. *J. Chem. Phys.* **2000**, *113*, 5154.

(49) Hattig, C.; Hald, K. *Phys. Chem. Chem. Phys.* **2002**, *4*, 2111.

(50) Wilson, A. K.; Almlöf, J. *Theor. Chem. Acc.* **1997**, *95*, 49.

(51) Hirata, S.; Head-Gordon, M. *Chem. Phys. Lett.* **1999**, *314*, 291.

(52) Onsager, L. *J. Am. Chem. Soc.* **1936**, *58*, 1486.

(53) Chipman, D. M. *J. Chem. Phys.* **2000**, *112*, 5558.

(54) Chipman, D. M. *Theor. Chem. Acc.* **2002**, *107*, 80.

(55) Chipman, D. M.; Dupuis, M. *Theor. Chem. Acc.* **2002**, *107*, 90.

(56) Shao, Y.; Molnar, L. F.; Jung, Y.; Kussmann, J.; Ochsenfeld, C.; Brown, S. T.; Gilbert, A. T. B.; Slipchenko, L. V.; Levchenko, S. V.; O'Neill, D. P.; DiStasio, R. A., Jr.; Lochan, R. C.; Wang, T.; Beran, G. J. O.; Besley, N. A.; Herbert, J. M.; Lin, C. Y.; Voorhis, T. V.; Chien, S. H.; Sodt, A.; Steele, R. P.; Rassolov, V. A.; Maslen, P. E.; Korambath, P. P.; Adamson, R. D.; Austin, B.; Baker, J.; Byrd, E. F. C.; Dachsel, H.; Doerksen, R. J.; Dreuw, A.; Dunietz, B. D.; Dutoi, A. D.; Furlani, T. R.; Gwaltney, S. R.; Heyden, A.; Hirata, S.; Hsu, C.-P.; Kedziora, G.; Khalliulin, R. Z.; Klunzinger, P.; Lee, A. M.; Lee, M. S.; Liang, W.; Lotan, I.; Nair, N.; Peters, B.; Proynov, E. I.; Pieniazek, P. A.; Rhee, Y. M.; Ritchie, J.; Rosta, E.; Sherrill, C. D.; Simmonett, A. C.; Subotnik, J. E.; Woodcock, H. L., III; Zhang, W.; Bell, A. T.; Chakraborty, A. K. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3172.

(57) Li, G.; Jiang, K.-J.; Bao, P.; Li, Y.-F.; Li, S.-L.; Yang, L.-M. *New J. Chem.* **2009**, *33*, 868.

(58) Dreuw, A.; Weisman, J. L.; Head-Gordon, M. *J. Chem. Phys.* **2003**, *119*, 2943.

(59) Dreuw, A.; Head-Gordon, M. *Chem. Rev.* **2005**, *105*, 4009.

(60) Sobolewski, A. L.; Domcke, W. *Chem. Phys.* **2003**, *294*, 73.

(61) Schreiber, M.; Silva-Junio, M. R.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2008**, *128*, 134110.

(62) Silva-Junio, M. R.; Schreiber, M.; Sauer, S. P. A.; Thiel, W. *J. Chem. Phys.* **2008**, *129*, 104103.

(63) Jacquemin, D.; Wathelet, V.; Perpète, E. A.; Adamo, C. *J. Chem. Theory Comput.* **2009**, *5*, 2420.

CT100069Q

JCTC Journal of Chemical Theory and Computation

# Accuracy of Computed $^{15}N$ Nuclear Magnetic Resonance Chemical Shifts

Adriana Gregušová, S. Ajith Perera, and Rodney J. Bartlett*

*Quantum Theory Project, Department of Chemistry and Physics, P.O. Box 118435, Gainesville, Florida 32611*

Received October 29, 2009

**Abstract:** Benchmark CCSD(T) $^{15}N$ NMR calculations are performed for 35 experimentally known $^{15}N$ shifts of 29 molecules. For the eight known gas phase experimental values of $N_2$, HCN, $CH_3CN$, $N\underline{N}O$, $NH_3$, $\underline{N}NO$, $(CH_3)_3N$, and $CH_3NH_2$, CCSD(T) with a basis set previously calibrated for $^{13}C$ shifts is accurate to 0.2−3 ppm except for the $\underline{N}NO$ shift, which shows a deviation of 6 ppm. However, the differences between the computed and experimental values in solution due to solvent and finite temperature effects can be as large as ~25 ppm and must be estimated to relate gas phase 0 K computed values to experiment. An empirical correction is obtained by studying the variations between the estimated solvent effects and the absolute shielding constant. It is shown that the average deviation of computed shifts falls to 3.6 ppm from 12.6 ppm when the correction is applied.

## Introduction

Gauge origin independent NMR shielding calculations[1−15] with electron correlation effects reported at the second order many body perturbation theory (MBPT(2))[16−18] have been extended to the predictive infinite order coupled cluster (CC) methods[18] including some triple excitation effects at various levels (CC single doubles and triples, CCSDT) and CCSD(T).[18−22] Also, multiconfiguration self-consistent field (MCSCF) methods have been applied for special cases with large nondynamic correlation effects.[23,24] Density functional theory (DFT) methods,[25−30] which are applicable for large molecules,[10−14] still suffer from accuracy and reliability issues which prohibit offering a truly predictive tool.

The accuracy of computed values, a prerequisite for validating experimental assignments or making new assignments, or in some cases supplanting the experiments, are customarily assessed by comparing the computed results either to experiment or to the full configuration interaction results. The latter gives an unambiguous measure of the errors associated with the theoretical approximations for *a given basis set* and geometry. However, the full CI NMR shielding calculations are only available for the $H_2$ molecule.[31] Comparisons with reliable experimental data are far more important to gauge the accuracy of the computed data. However, these are complicated by the fact that the theory cannot account for the solution and the other conditions of most experiments.

In practice, a judicious choice has to be made for what experimental data is best used for the property under consideration, since the experiments are done in conditions that are far from ideal (noninteracting isolated molecules at 0 K) as assumed in calculations. For example, shielding constants are known to be highly influenced by the nature of the solvent effects.[32−35]

Another matter of concern in comparisons with experiment is the inability to choose the geometry and the basis sets. For example, which geometry, optimized, experimental, or in solution, is the most appropriate. Since there are no clear-cut choices, this step is best done by using a series of basis sets and geometries to establish the accuracy of the method compared to experiment for a given geometry and basis set.

Gauss and co-workers presented a systematic study of the accuracy of the $^{13}C$ NMR shielding constants computed by CC methods.[36] The results from this paper and from a series of papers published earlier by Gauss, Stanton, and co-workers[4,10−12,14,16−22] form the basis for the current knowledge of the accuracy of computed NMR shielding constants for the widely used NMR isotopes by CC and MBPT methods. In this paper our focus is on the accuracy of the computed $^{15}N$ NMR shielding constants in solutions.

* Corresponding author e-mail: bartlett@qtp.ufl.edu.

10.1021/ct9005739 © 2010 American Chemical Society
Published on Web 03/04/2010

In the context of the accuracy of computed $^{15}N$ chemical shifts in the present work, two recent papers are of great interest. One of them is the widely publicized work on the synthesis and the identification of the $N_5^+$ molecule,[37] the only other homoleptic polynitrogen species that is known beside $N_2$ and the azide anion ($N_3^-$). Among the spectroscopic techniques used, the $^{15}N$ NMR is regarded as the most definitive, and the agreement between the computed chemical shifts for the $N_5^+$ molecule at its optimized geometry and the measured shifts is quoted as the strongest evidence to support the presence of $N_5^+$. The computed values are obtained at the CCSD(T) level by Stanton[38] and are expected to be within 1−2 ppm from the gas phase measurements analogous to the accuracy of the computed values of $^1H$ and $^{13}C$ NMR chemical shifts. However, the experiments are done in highly polar solvents, and only after an empirical correction is applied to account for the condensed phase effects can the agreement between scaled and measured values be considered adequate to support the conclusion that the observed NMR corresponds to $N_5^+$. While this is a successful application of computational methodology, we note that the correction for the condensed phase effects is chosen with severely inadequate numerical support and the accuracy of the CCSD(T) $^{15}N$ shifts for the geometry and basis set used needs further assessment.

The recent report of the potential isolation of the long predicted[39−41] pentazole anion ($N_5^-$) in solution by Butler et al.[42] is another example that employs the computed $^{15}N$ NMR chemical shifts to assign the measured spectra. In this case, the computations use DFT levels of theory. Despite large error bars (±20 ppm), a match between the computed and measured shifts is one of the two pieces of evidence quoted as indicating the presence of $N_5^-$. However, Schroer et al.[43] who attempted to independently verify these findings by replicating the experiments of Butler et al.[42] argue that the $^{15}N$ NMR signal that Butler et al.[42] assigned to the nitrogen atoms in $N_5^-$ is actually due to the nitrogen atom in $NO_3^-$ which is abundant in the reaction mixture. Hence, Schroer et al.[43] question the validity of the claim that the $N_5^-$ had been observed. Since the accuracy of the DFT level NMR chemical shifts are known to vary from case to case, a match between the DFT values with the measured ones may not be used as prima facie evidence to establish the presence of $N_5^-$. We also recognize that Schroer et al.'s arguments are exclusively based on the published NMR shifts of the primary species present in the experiment and are not further supported by computations. Part of the resolution of these conflicting experiments rests upon having reliable and also highly accurate $^{15}N$ NMR calculations of all the primary species involved in the experiment, in solution. We have pursued such a study, and the details are presented in a separate paper[44] devoted to the theoretical evaluation of the experimental findings of the Butler et al. and Schroer et al.[42,43] experiments and the findings in another follow up paper by Butler et al.[67] In that paper we also address the scalar spin−spin coupling constants.

In this study our focus is on the accuracy of the computed $^{15}N$ NMR shielding constants. When reliable gas phase experimental $^{15}N$ shielding data are available, analysis similar to the $^{13}C$ work of Gauss and co-workers for $^{13}C$ NMR can be performed to augment the computed data set to include the $^{15}N$ results. However, practical problems such as those involving $N_5^+$ and $N_5^-$ described above raise additional concerns that need to be addressed. First and foremost experiments are done in solution (most cases highly polar), and a practically viable, justifiable scheme to obtain a measure of the solvent effects must be established. Moreover, since it has been noted that the solvent effects are more pronounced for $^{15}N$ shielding than those for $^{13}C$, correcting for the solvent effects is even more relevant in the case of $^{15}N$ NMR chemical shielding calculations.[45] It is impractical to assume that we can incorporate the solvent effects on a first principle basis in the reasonable future for complicated solutions, though certainly electrostatic cavity models could be used,[46,47] as well as classical water force fields[48,49] for some cases. Lacking this capability, we advocate looking at the differences between the measured shifts in any solution and the very accurately computed gas phase values for a series of molecules to possibly identify systematic variations.

Another issue that needs attention is the molecular size, since it is not always possible to do a large basis set CCSD(T) calculation for the individual molecules that are of interest in experiments. For example, in the case of the above experiments, the precursors and the decomposition products in the synthesis are also $^{15}N$ NMR active, and they are larger molecules than the $N_5^-$ itself; a full resolution of the measured spectra might require knowledge of their NMR as well. It is now generally accepted that electron correlation effects must be included, preferably at least at the CCSD(T) level with large basis sets, but when such high level calculations are impractical, at a minimum a small basis MBPT(2) (or DFT) result at the best possible geometry (preferably experimental) must be obtained (minimum threshold). The difference between the minimum threshold and the best possible calculations (CCSD(T) with large basis sets) must be known with some certainty.

While mindful of these special requirements for the problem we plan to address, the following concerns are significant: (1) We consider the accuracy of the $^{15}N$ CCSD(T)/pz3d2f shielding constants when gas phase experimental data is available and the corresponding CCSD(T) calculations that can be performed. The geometries are optimized at the CCSD(T) level using cc-pVQZ or aug-cc-pVQZ basis sets. (2) We assess the variation of the difference between the MBPT(2) and CCSD(T) $^{15}N$ shieldings for a given geometry with respect to basis set choice. We want to explore the possibility of estimating CCSD(T)/pz3d2f values for larger systems for which such calculations are impractical by first computing the MBPT(2)/pz3d2f results followed by adding the difference between CCSD(T) and MBPT(2) shieldings computed using a much smaller basis set such as cc-pVTZ. In order to be successful the difference between the CCSD(T) and MBPT(2) shielding must be insensitive to basis set. If successful, this will expand the range of molecules that we can access since there are many cases where the MBPT(2)/pz3d2f calculations are practical but not those for CCSD(T)/pz3d2f. We use cc-pVTZ or cc-pVDZ optimized geometries for larger molecules. (3) We consider

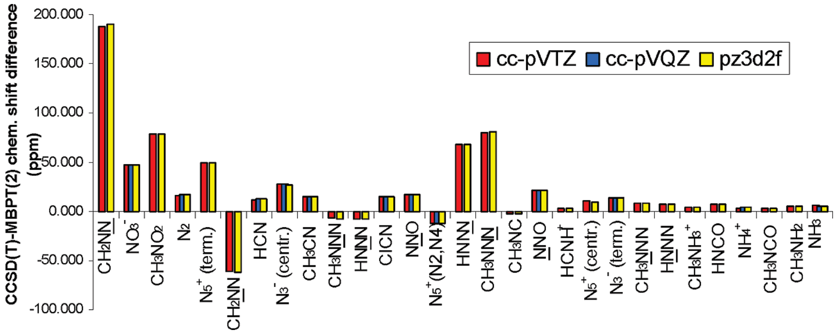


***Figure 1.*** Differences between CCSD(T) and MBPT(2) calculated $^{15}N$ NMR chemical shifts (ppm). Different colors represent various basis set results for the molecules from set-A.

the difference between the computed or estimated CCSD(T)/pz3d2f and the condensed phase experimental results to quantify our averaged condensed phase effects and possibly identify correlations.

The details of all the NMR calculations and geometry optimizations, including a description of various basis sets used, the calculation level, and the molecules considered, are presented in the Theoretical Calculations section. It is followed by sections devoted to results, their analysis, and conclusions.

## Theoretical Calculations

Both the ACES II UF[50] and ACES II MAB[51] versions of the ACES II program system are used for the calculations. The geometries are optimized at the CCSD(T) level employing Dunning et al.[52] correlation consistent polarized valence and augmented basis sets: CCSD(T)/aug-cc-pVQZ level optimizations for $N_5^-$ and $N_2O$; CCSD(T)/cc-pVQZ level for $N_3^-$, $N_5^+$ $NO_3^-$, $NH_3$, $NH_4^+$, $N_2$, HCN, $HCNH^+$, $CH_2NN$, $CH_3CN$, and ClCN; CCSD(T)/cc-pVTZ level for $CH_3NO_2$, $CH_3NC$, HNCO, $CH_3NCO$, $CH_3CNH^+$, $CH_3NH_2$, $CH_3NH_3^+$, $CH_3NNN$, $HN_3$, and $N_5H$; and CCSD(T)/cc-pVDZ level for the others. All the CCSD(T)/cc-pVQZ and CCSD(T)/aug-cc-pVQZ optimized geometries are confirmed to be minima by computing the harmonic frequencies. The $^{15}N$ NMR chemical shielding constants are computed at the MBPT(2) and CCSD(T) levels by employing gauge-including atomic orbitals (GIAOs). The basis sets, pz3d2f (pz3p2d for hydrogen), cc-pVTZ, and cc-pVQZ, are selected. As discussed earlier, the pz3d2f and pz3p2d (H) basis sets are shown to provide $^{13}C$ NMR chemical shielding in close agreement with experimental results for a set of 16 molecules (standard deviation of CCSD(T)/pz3d2f shielding from experiment is 1.3 ppm)[36] and are chosen as our benchmark basis sets. The pz3d2f and pz3p2d (H) basis sets are derived from the original basis set of Ahlrichs and co-workers[53] using polarization function exponents from ref 54. The two other sets, cc-pVTZ and cc-pVQZ, are used whenever the pz3d2f and pz3p2d (H) series are beyond the level of available computer resources. Nitromethane is used as the external reference (experimental absolute shielding) to obtain the $^{15}N$ NMR chemical shifts. We have also used $NH_3$ as an internal standard (absolute shielding at CCSD(T)/pz3d2f level) for those molecules for which the gas phase experimental results are available.

## Results and Discussions

Before proceeding to present our results, let us briefly summarize the points discussed previously about the nature of various calculations and their purpose and the organization of the results. Our primary target is to compute CCSD(T)/pz3d2f quality $^{15}N$ chemicals shifts for a series of nitrogen containing molecules. For the cases where such direct calculations are impractical, they are obtained indirectly. The molecules whose results can be obtained directly at the CCSD(T)/pz3d2f level are labeled as set-A. Similarly, set-B consists of molecules for which the highest level NMR calculations currently possible are the CCSD(T)/cc-pVTZ and MBPT(2)/pz3d2f, and set-C consists of molecules for which the highest level NMR calculations currently possible are the MBPT(2)/cc-pVDZ, MBPT(2)/cc-pVTZ, and CCSD(T)/cc-pVDZ. As a consequence, for those molecules in both set-B and set-C, the CCSD(T)/pz3d2f quality values must be estimated. The individual molecules in each category follow: set-A, $CH_2NN$, $NO_3^-$, $CH_3NO_2$, $N_2$, $N_5^+$, HCN, $N_3^-$, $CH_3CN$, $CH_3NNN$, HNNN, ClCN, NNO, $CH_3NC$, $HCNH^+$, $CH_3NH_3^+$, HNCO, $NH_4^+$, $CH_3NCO$, $CH_3NH_2$, $NH_3$, $N_5^-$, and $N_5H$; set-B, $C_3H_4N_2$ (12-diazole), $C_4H_4NH$ (pyrrole), $CH_3CNH^+$, $(CH_3)_3N$, and $CH_3CH_2NH_2$; set-C, $MeOC_6H_4N_5$ and $MeOC_6H_4N_3$.

The first series of data presented is to establish the proposed scheme to estimate the CCSD(T)/pz3d2f quality results for the systems that are not directly amenable to the CCSD(T)/pz3d2f calculations, i.e., molecules in set-B and set-C. Then, a comparison of our calculated values with the gas phase experimental values will be presented. Finally, the computed results are compared with the available experimental data irrespective of the nature of the medium of the experiments. The purpose of this comparison is to see whether a justifiable empirical correction can be established that in general can apply to the computed values to account for the medium effects which are absent in the theoretical models that we have used to make valid comparisons with the condensed phase experiments.

**Variation of the Difference between CCSD(T) and MBPT(2) Shielding Constants with Basis Sets.** Figure 1 shows the difference of the CCSD(T) and the MBPT(2) $^{15}N$ shielding constants for nitrogen atoms in 28 molecules

***Table 1.*** Comparison of the Gas and Liquid Phase Experimental $^{15}N$ NMR Chemical Shifts (ppm) with Those Calculated at CCSD(T)/pz3d2f Level

| nitrogen position | $N_2$ | HCN | $CH_3CN$ | $N\underline{N}O$ | $\underline{N}NO$ | $(CH_3)_3N$ | $CH_3NH_2$ | $NH_3$ |
|---|---|---|---|---|---|---|---|---|
| gas phase expt chem shift (ppm) | −75.3[a] | −115.4[b] | −126.7[a] | −148.0[a] | −232.3[a] | −372.8[c] | −385.4[c] | −400.1[a] |
| liquid phase expt chem shift (ppm) | −70.2[a] | −127.5 | −137.1[d] | −140.0[c] | −225.0[d] | −363.1[d] | −377.3[d] | −380.2[d] |
| CCSD(T)/pz3d2f chem shift (ppm) | −77.5 | −123.0 | −132.3 | −149.9 | −244.9 | −370.0[e] | −392.8 | −407.2 |
| CCSD(T)/pz3d2f chem shift (ppm) with $NH_3$ as internal standard | −70.7 | −116.1 | −125.4 | −143.1 | −238.1 | −363.1 | −385.9 | −400.3 |

[a] Reference 57. [b] Reference 58. [c] Reference 60. [d] Reference 59. [e] CCSD(T)/pz3d2f chem shift for $(CH_3)_3N$ was estimated using MBPT(2)/pz3d2f, CCSD(T)/cc-pVTZ, and MBPT(2)/cc-pVTZ results, and the constant basis set correction (see detailed explanation in the text for set-B molecules).

obtained with the cc-pVTZ, cc-pVQZ, and pz3d2f basis sets. Note that most of the molecules in this set have multiple nitrogen atoms and hence the corresponding multiple shielding constants. For convenience, unless it becomes necessary within the context of the discussion, no reference to individual nitrogen atoms in a given molecule is made.

As we can see, the difference between CCSD(T) and MBPT(2) shielding constants is nearly invariant with respect to basis sets for a given nitrogen atom. Also, we note that this difference can be as large as 200 ppm ($CH_2NN$) or as small as 1−2 ppm while the average is 18 ppm (the average is computed using CCSD(T)/cc-pVTZ results[36]). Another observation is that the difference is largest for NN−X, where X = C, N, O, etc., systems. Further analysis of the results for the NN−X type systems included in this selection of molecules shows that the largest difference occurs for the terminal nitrogen atom, and it is positive while for the central nitrogen the error is smaller (compared to that of the terminal nitrogen) and often negative. The association of the relative change in the density to the shielding constant is the basis for all the empirical rules which have been used very successfully in relating the structures to the corresponding NMR spectra.[55,56] While it is an interesting point to further investigate the nature of the larger difference between CCSD(T) and MBPT(2) for certain bonding situations, that is left for a future study.

The pertinent observation here is that the MBPT(2) and CCSD(T) shielding constant differences are insensitive to basis set, since that provides a mechanism to reliably estimate the CCSD(T)/pz3d2f results for the molecules in set-B and set-C (for molecules in set-A, the CCSD(T)/pz3d2f level shielding constants can be computed directly). For the molecules in set-B, we first compute the MBPT(2) and CCSD(T) shielding constants with the cc-pVTZ basis set and the MBPT(2) shielding constants with the pz3d2f basis set. The difference between MBPT(2) and CCSD(T) computed with the cc-pVTZ basis set is used to correct the computed MBPT(2)/pz3d2f basis set results to obtain an estimate for the CCSD(T)/pz3d2f results using the fact that the difference is relatively insensitive to the basis sets. For the molecules in set-C, we can directly compute shielding constants at the CCSD(T)/cc-pVDZ, MBPT(2)/cc-pVDZ, and MBPT(2)/cc-pVTZ levels. The basis set independence of the shielding constant difference between the CCSD(T)/cc-pVDZ and MBPT(2)/cc-pVDZ results is used to scale the MBPT(2)/cc-pVTZ values to obtain an estimate for the CCSD(T)/cc-pVTZ shielding constants.

**Comparison of the CCSD(T)/pz3d2f Results with Gas Phase Experimental Results.** As we have pointed out earlier, it is best to compare computed shielding constants (or shifts) to gas phase experimental shielding constants (or shifts). To the best of our knowledge there are only eight gas phase experimental $^{15}N$ NMR chemical shifts reported in the literature. They are shown in Table 1 along with the computed values at the CCSD(T)/pz3d2f level and the corresponding liquid phase experimental results. The CCSD(T)/pz3d2f chemical shifts using $NH_3$ as an internal standard are also included in Table 1 for comparison. It can be seen that, with respect to $NH_3$, internal standard computed shifts for polar N atoms are in closer agreement with the gas phase experiment than with respect to the $CH_3NO_2$ external standard. However, for less polar or nonpolar N atoms, the gas phase experimental and theoretical chemical shift differences is larger. Therefore, while discussing the results in Table 1, we use chemical shifts with $NH_3$ as an internal standard for polar molecules and $CH_3NO_2$ as an external standard for N atoms in less polar or nonpolar molecules. The absolute deviations of such defined CCSD(T)/pz3d2f computed values are within 0.2−3 ppm of the gas phase experimental results, which is slightly higher than the Gauss and co-workers findings for the $^{13}C$ NMR shielding constants using the same level of theory.[36] The biggest deviation (6 ppm) occurs for the terminal nitrogen atom of NNO. This may be an indication that the terminal nitrogen in NN−X type systems may need correlation effects that go beyond the CCSD(T) level. We note that the CCSDT/qz2p calculations of NMR chemical shifts for NNO reported by Gauss[61] are in very good agreement with our CCSD(T)/pz3d2f results (they are within 0.2 and 0.6 ppm for terminal and central nitrogen atom in NNO, respectively).

As discussed previously, we have also noted that the terminal nitrogen in other NN−X systems shows the largest difference between CCSD(T) and MBPT(2) results irrespective of the basis set size. It would be interesting to further evaluate the special nature of the terminal nitrogen in the NN−X type systems.

Note that the gas phase experiments are conducted at finite temperatures and the molecules can vibrate and rotate freely. In this work, we have not considered the ro-vibrational effects or the relativistic effects. The prior work done with the $^{13}C$ NMR shielding constants reported that on the average ro-vibrational effects amount to 1−2 ppm.

**Comparison of the CCSD(T)/pz3d2f Results with Experimental Results: Set-A and Set-B.** Most of the $^{15}$N NMR experimental values in the literature are obtained in liquid phase at finite temperatures, far from the ideal conditions that we assume in computations. Hence, comparisons of computed values with experiments are meaningful only when the solvent and finite temperature effects are incorporated. Furthermore, it is important to point out that not having a uniform set of well established experimental data due to the fact that different experiments use different conditions (different concentrations of solutions, temperature, etc.) makes comparisons with experiment even more difficult.

The experimental $^{15}$N data for 28 molecules considered here are published by three different groups. The experimental values along with the conditions under which they are obtained are presented in Table 2. For the most part the results of Jameson and co-workers are obtained in neat liquids[57] while the results of Levy and co-workers[45] and Berger and co-workers[59] are obtained mainly in solution. A plot of the difference of CCSD(T)/pz3d2f computed values (for the molecules in the set-B CCSD(T)/pz3d2f values are estimated according to the schemes presented earlier) compared to the experimental values is shown in Figure 2. In contrast to the uniform 0.2−3 ppm deviation of computed results with the gas phase experiment, deviation of the computed values from the measured values of experiments in solution shows a wide variation. For example, in some cases it is as small as 1 ppm and in others it is as large as 28 ppm and can be either positive or negative. The mean absolute deviation from experiment is 12.6 ppm while the mean absolute deviations of the positive and negative chemical shielding taken separately are 15.9 and 5.2 ppm, respectively. As discussed previously, besides a couple of ppm error due to vibrational and relativistic effects, the remaining difference between the computed and measured results is entirely due to the medium effects and the residual correlation and basis set effects that are absent in the current level of theory. As a result of the variations in the measured chemical NMR shifts depending upon the experimental conditions, there are multiple values for the difference between computed (gas phase) and measured values for a given atom in a molecule. For example, in the case of $NO_3^-$, the measured chemical shifts vary from −12.6 to −4.0 ppm depending on the experimental conditions (a likely cause is the acidity of the medium). It is quite clear from these results that it is ill advised to directly use the computed gas phase chemical shifts to assist assignments of experiments in solution. Since there are no well established first principle based methods to incorporate solvent effects on NMR chemical shift calculations, we focus on devising an empirical correction as an alternative. We note that the success of such an empirical correction depends on whether the solvent effects can be approximately treated as an average rather than specific to individual experiments.

In the process of calibration of a new method, a statistical analysis of the differences between the computed and measured values is used to establish the accuracy and the expected error bars. For this purpose the experimental values are carefully chosen so that the differences between the two values are small and only due to the deficiencies in the theoretical treatment of the interactions in the entire experiment rather than due to the neglect of them. But in practice, this is not always possible and when what is being neglected in the computation is large, as is the case of the solvent and finite temperature effects in this work, one can expect large and less systematic deviations. It is not uncommon in the literature that the average differences are used as a measure of the effects that are untreated (or approximated). However, it is important to emphasize that the empirical corrections such as the ones described below must be established with large data sets and should be known with respect to a quantifiable parameter that is directly influenced by the effects that are neglected in the calculation.

Strictly speaking the shielding constant difference between the experimentally measured shielding constants in the gas phase and in a solvent give a quantitative measure of that solvent effect. The measured gas phase values are often unavailable, but they can be accurately computed. Assuming that the computed values are converged to the measured gas phase values, the difference between the measured and computed results gives an estimate of the solvent effects. A plot of the difference of computed and measured chemical shifts obtained under a variety of experimental conditions against the computed absolute shielding constants of the corresponding nuclei is shown in Figure 3.

It is pertinent to note several interesting trends that can be observed in the data presented in this figure. One obvious and expected trend is the increase of the absolute difference between the computed and measured shifts as the absolute shielding constant increases. Also, within the data set that we have considered, for the positive shielding constants, the difference is also positive, and for the negative shielding constants, in contrast, the difference is either positive or negative. Furthermore, for the negative shielding constants, the difference is much smaller in comparison with that for the positive shielding constants.

NMR shielding constants are a measure of the response to an external magnetic field ($B_0$) due to a change in the electron density caused by bonding (relative to the bare nuclei) in the vicinity of the nuclei of interest (the effective magnetic field at the nucleus, $B_{\text{eff}}$, is then $B_{\text{eff}} = B_0 + \sigma B_0$ where $\sigma$ may be positive or negative). We argue that, for all practical purposes, on average the solvent acts as a perturbation of the electron density; consequently, the magnitudes of the shielding constants due to the solvent effects, estimated by the difference between the computed and the measured shielding constants, are expected to show correlation with these computed shielding constants. Let us denote the shielding constants of an isolated atom $\sigma_0$ and its corresponding value when it is bonded as $\sigma$, respectively. Also, denote the changes in the shielding constants in the presence of an external medium such as a solvent as $\Delta\sigma_0$ and $\Delta\sigma$, respectively (we note that the $\Delta\sigma_0$ is not measured under normal circumstances). The shielding constant $\sigma_s$ in the presence of a solvent can then be expressed as $\sigma_s \propto (\sigma + \Delta\sigma) - (\sigma_0 + \Delta\sigma_0)$. A more useful form of this expression for discussion purposes is $\sigma_s \propto (\sigma - \sigma_0) + (\Delta\sigma - \Delta\sigma_0)$, where $\sigma_g = (\sigma - \sigma_0)$ and $\Delta\sigma_s = (\Delta\sigma - \Delta\sigma_0)$ are to a good

***Table 2.*** Summary of the Reported Experimentally Measured $^{15}N$ NMR Chemical Shifts (ppm) and Conditions of Their Measurement

| chem shift (ppm)/conditions | expt gas phase | expt liquid phase | expt solution (solvent) | Schroer | Butler |
|---|---|---|---|---|---|
| $CH_2N\underline{N}$ | | | 16.8 ($CDCl_3$)[b] | | |
| | | | 7.8 ($Et_2O$)[c] | | |
| $NO_3^-$ | | | −4.0 (satd $NH_4NO_3$)[a] | −11.5 | |
| | | | −4.6 (5 M $NH_4NO_3$,2 M $HNO_3$)[a] | | |
| | | | −12.6 (7 M $HNO_3$)[a] | | |
| | | | −3.7 (satd $NaNO_3$)[b] | | |
| $CH_3NO_2$ | | 0.0 | | | |
| 1$\underline{2}$diazole | | | −182.0 (av, $CH_3OH$)[b] | | |
| | | | −134.7 (av, $CHCl_3$)[b] | | |
| | | | −134.0 (av, $CDCl_3$)[c] | | |
| | | | −79.8 ($(CH_3)_2SO$)[c] | | |
| $N_2$ | −75.3 (303 K)[a] | −74, −70.2 (77 K)[a] | −70.1 (2 M $HNO_3$)[b] | | −72.0 |
| $N_5^+$ (term.) | | | | −100.4[e] | |
| $CH_2\underline{N}N$ | | | −94.2 ($CDCl_3$)[b] | | |
| | | | −96.0 ($Et_2O$)[c] | | |
| HCN | | −127.5 (300 K)[a], | | | |
| | | −129 (309 K)[a] | | | |
| $N_3^-$ (centr) | | | −130.4 ($NaN_3$ in $H_2O$)[c] | −133.7 | −147 |
| $CH_3CN$ | −126.7 (227 K)[a] | −138.0 (227 K)[a], | −140.7[b] | | |
| | | −136.4(303K)[a] | | | |
| | | −137.1[c] | | | |
| $CH_3N\underline{N}N$ | | | −133.2 ($CDCl_3$)[b] | | |
| | | | −129.7 (benzene)[c] | | |
| $HN\underline{N}N$ | | | −134.5 (ether)[c] | | |
| ClCN | | −144 (309 K)[a] | | | |
| $N\underline{N}O$ | −148 ($CCl_4$, 303 K)[a] | −142 (193 K)[a] | −140.0 (EBBA[d])[c] | | |
| $N_5^+$ (N2, N4) | | | | −165.3[e] | |
| $HNN\underline{N}$ | | | −179.0 (ether)[c] | | |
| $CH3NN\underline{N}$ | | | −171.0 (benzene)[c] | | |
| $\underline{1}$2diazole | | | −182.0 (av, $CH_3OH$)[b] | | |
| | | | −134.7 (av, $CHCl_3$)[b] | | |
| | | | −134.0 (av, $CDCl_3$)[b] | | |
| | | | −173.1 ($(CH_3)_2SO$)[c] | | |
| $CH_3NC$ | | −218.2[b] | | | |
| | | −218.0[c] | | | |
| $\underline{N}NO$ | −232.3 ($CCl_4$, 303 K)[a] | −226 (193 K)[a] | −225.0 (EBBA[d])[c] | | |
| pyrrole | | −231.4[b] | −236.4 ($CCl_4$)[b] | | |
| | | −231.4[c] | −222.3 ($(CH_3)_2SO$)[b] | | |
| $HCNH+$ | | | −235.8 ($FSO_3H-SbF_5-SO_2$)[b] | | |
| $N_5^+$ (centr) | | | | −237.3[e] | |
| $CH_3CNH^+$ | | | −239.6 ($FSO_3H-SbF_5-SO_2$)[b] | | |
| | | | −239.2 ($FSO_3H-SbF_5-SO_2$)[c] | | |
| $N_3^-$ (term.) | | | −280.9 ($NaN_3$ in $H_2O$)[c] | −281.5 | −283 |
| $CH_3\underline{N}NN$ | | | −321.2 (benzene)[c] | | |
| $H\underline{N}NN$ | | | −324.9 (ether)[c] | | |
| $(CH_3)_3N$ | −372.8[f] | −367.2[b] | −363.1 ($CH_3OH$)[c] | | |
| $CH_3NH_3^+$ | | | −361.4 (1 M in $CH_3OH$)[b] | | |
| HNCO | | | −346.0 ($C_6H_{12}$)[c] | | |
| $CH_3CH_2NH_2$ | | | −355.4 ($CH_3OH$, 273 K) | | |
| $NH_4^+$ | | | −359.6 (satd $NH_4NO_3$)[a] | −359.8 | |
| | | | | −361.3 | |
| | | | | −362.0 | |
| | | | | −362.7 | |
| $CH_3NCO$ | | −366.1[b] | | | |
| | | −365.3[c] | | | |
| $CH_3NH_2$ | −385.4[f] | −378.9[b] | −377.3 ($CH_3OH$, 273 K)[c] | | |
| $NH_3$ | −400.1 (195 K)[a] | −377.5 (195 K)[a] | | | |
| | −399.3 (302 K)[a] | −382.1 (303 K)[a] | | | |
| | −396.1 (5 atm)[b] | −381.9 (303 K)[a] | | | |
| | | −380.4 (300 K)[a] | | | |
| | | −380.2 (298 K)[a] | | | |
| | | −380.2 (298 K)[b] | | | |
| | | −376.9 (223 K)[b] | | | |

[a] Reference 57. [b] Reference 45. [c] Reference 59. [d] EBBA: *N*-(*p*-ethoxybenzylidene)-*p*-*n*-butylaniline. [e] Reference 37. [f] Reference 60.

approximation a measure of the gas phase shielding constant and the solvent effects, respectively. Thus, when the computed gas phase ($\sigma_g$) shielding constant is positive and large, then the solvent effect, ($\Delta\rho - \Delta\rho_0$), is most likely to be

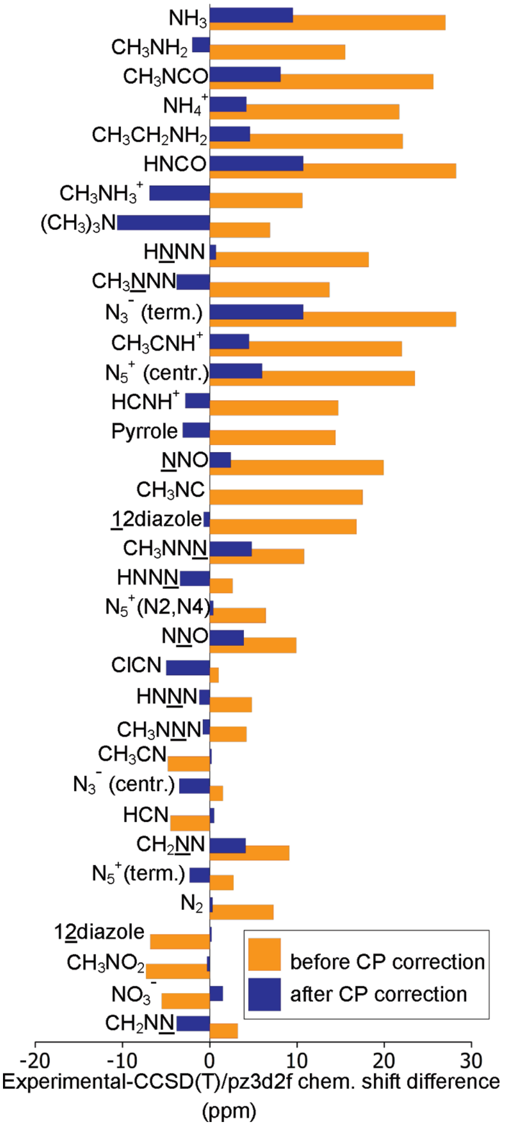


***Figure 2.*** Differences of experimental $^{15}N$ NMR chemical shifts (ppm) and those calculated at the CCSD(T)/pz3d2f level with and without the empirical correction for solvent effects.

large and positive. However, when $\sigma_g$ is negative, the solvent effect can be either slightly positive or negative.

A closer look at Figure 3 shows that we can identify regions of shielding constants where on average the contribution due to the medium remains unchanged. For example, we can see that, for shielding constant values in the range 0−50 ppm, the average difference is 6 ppm with the minimum and maximum differences being 1 ppm and 11 ppm, respectively. Similarly, for shielding constants in the range >50 ppm, the average is 17 ppm while the maximum and minimum differences are 7 and 28 ppm, respectively. For negative shielding constants, we can identify that the 0−(−)50 ppm range has an average error of ±5 ppm with maximum deviations of ±5 ppm and the range >(−)50 ppm has an average error of ±7 ppm while maximum and minimum deviations are ±7, respectively. These results are summarized in Tables 3 and 4.

If we use the average differences for each range listed above as the correction for the solvent effects and re-evaluate the differences between the experiment and the computed results, after applying the proposed corrections, we obtain the results shown in Figure 2 and Table 5. We note that the mean deviation (absolute values) from the experiment of the computed gas phase results decreases from 12.6 to 3.6 ppm. Also, note that after applying the correction the mean deviations of the positive and the negative shielding constants are 4.6 and 1.6 ppm, respectively, compared with the corresponding averages prior to applying the correction, 15.9 and 5.2 ppm.

The mean deviations from experiment for the molecules in set-A and set-B are 12.1 and 14.8 ppm, respectively. The mean deviation for the group of positive chemical shifts is 15.8 ppm (set-A) and 16.4 ppm (set-B, and that for the group of negative chemical shifts is 5.0 ppm (set-A) and 6.8 ppm (set-B)). These results and the results presented in the prior section point to a general pattern that the deviation from experiment for positive shielding constant is comparatively larger than that for negative shielding constants. Comparing the behavior of sets-A and -B shows that mean deviations for set-B are only slightly higher. This might be caused by the error in the estimate used to obtain the CCSD(T)/pz3d2f chemical shift values in set-B as well as by the size of the statistical sample, since set-B only contains 6 values (5 corresponding to the positive chemical shifts and only 1 corresponding to the negative chemical shift).

**Comparison of the CCSD(T)/cc-pVTZ Results with Experimental Results: Set-C.** The molecules considered in set-C, further subdivided into subsets C1 and C2, having 18 or 20 atoms are the largest molecules presented in this work. As a consequence, the quality of the calculations (basis sets and theoretical method) is severely hampered. In Table 6 are presented computed $^{15}N$ NMR data for the molecules in set-C. They include MBPT(2) and CCSD(T) results obtained with the cc-pVDZ basis set and the MBPT(2) results obtained with cc-pVTZ, and these are currently the best possible calculations. The first observation is that the cc-pVDZ basis set, regardless of the level of theory, performs very poorly and is unsuitable for comparison with experiment. Nevertheless, the differences between the CCSD(T) and MBPT(2) obtained with the cc-pVDZ basis set are reliable and can be used to correct the MBPT(2)/cc-pVTZ results to estimate the CCSD(T)/cc-pVTZ values. These CCSD(T)/cc-pVTZ estimated values differ from 4 experimentally known values by 5.9, 11.1, 0.3, and 8.3 ppm, respectively, and, after correction for the effects of the medium, change to 0.9, 5.1, 7.3, and 1.3 ppm.

**Comparison of the CCSD(T)/pz3d2f Results with Previous Theoretical Results.** We have noted only a very few previous theoretical studies on $^{15}N$ NMR shifts, and the results from those studies along with the best values from this work and from experiment are shown in Table 7. It is evident from the results shown in Table 7 that MBPT(2) performs poorly when compared with other methods. Furthermore, the Mulliken populations which correlate with the NMR shielding indicate that the electronic structure of these molecules is inadequately described by MBPT(2). This leads us to conclude that the correlation effects beyond the MBPT(2) level must be included not only for the NMR but also for the other properties. The DFT-based methods tend

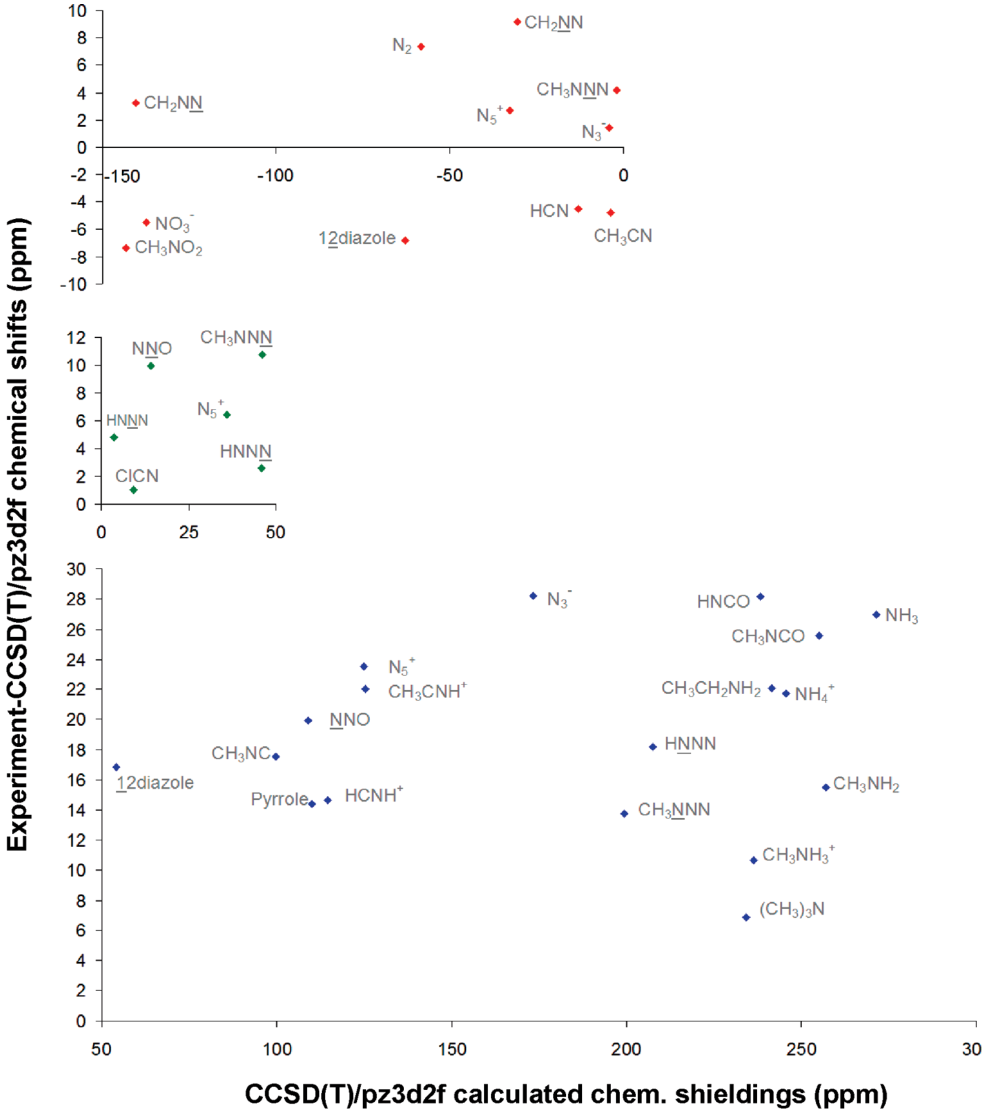


***Figure 3.*** Relationship of the $^{15}$N NMR chemical shift differences (experimental − CCSD(T)/pz3d2f, ppm) and the CCSD(T)/pz3d2f chemical shielding values (ppm).

***Table 3.*** CP Correction to CCSD(T)/pz3d2f $^{15}$N NMR Chemical Shifts for Atoms with Positive Absolute Shielding[a]

| absolute chem shielding value (ppm) | 0−50 | >50 |
|---|---|---|
| chem shift error − lower bound (ppm) | 1 | 7 |
| chem shift error − upper bound (ppm) | 11 | 28 |
| CP correction used − average (ppm) | 6 | 17.5 |

[a] Computed chemical shifts underestimate the experimental value.

***Table 4.*** CP Correction to CCSD(T)/pz3d2f Calculated $^{15}$N NMR Chemical Shifts for Atoms with Negative Absolute Shielding[a]

| absolute chem shielding value (ppm) | 0−(−)50 | >(−)50 |
|---|---|---|
| chem shift error − lower bound (ppm) | −5 | −7 |
| chem shift error − upper bound (ppm) | 5 | 7 |
| CP correction used (ppm) | 5 for positive error<br>−5 for negative error | 7 for positive error<br>−7 for negative error |

[a] If the N atom is internally polarized, the error is negative as in HCN. If the N atom is neutral, the error is positive as in $N_2$.

to do much better, and with the exception of a small number of DFT functional/molecule combinations, their results are consistent with the higher level computational methods. The CCSD(T)/pz3d2f results and CCSDT/qz2p chemical shifts calculated by Gauss[61] are in close agreement (the differences are 0.2, 0.6, and 2.2 ppm for terminal and central N atom in NNO and for $N_2$, respectively). Both CCSDT and CCSD(T) NMR shifts are in good agreement with the experimental shifts. In order to be consistent with the literature values, the CCSD(T) results in Table 7 are not corrected for solvent effects. As noted earlier, the experiments are done in solution or in some cases neat liquids, and direct comparison of experimental results with the gas phase values is unjustified. Nonetheless, the experimental data are also included in Table 7 for the purpose of having a basis for assessing the quality of different theoretical methods.

In the interest of “predictive” quantum chemistry it would be wonderful if solution effects on a solvated molecule could be described with the same accuracy as modern quantum chemistry can describe the isolated gas phase molecules.

***Table 5.*** Summary of $^{15}N$ NMR Chemical Shifts (ppm) Calculated at CCSD(T)/pz3d2f Level and CP Corrected and Their Comparison with Reported Experimental Values for the Molecules in Set-A and Set-B

| chem shift (ppm) | Schroer | Butler | most recent exptl value[e] | calcd CCSD(T)/pz3d2f | CP scaling factor | CP corrected chem shift |
|---|---|---|---|---|---|---|
| $CH_2N\underline{N}$ | | | 7.8 | 4.6 | 7 | 11.6 |
| $N_5H$ (N3) | | | | 3.2 | 7 | 10.2 |
| $NO_3^-$ | −11.5 | | −4.0 | 1.5 | −7 | −5.5 |
| $CH_3NO_2$ | | | 0.0 | 7.3 | −7 | 0.3 |
| $N_5^-$ | | | | −13.5 | 7 | −6.5 |
| $N_5H$ (N2) | | | | −34.6 | 7 | −27.6 |
| 1$\underline{2}$diazole | | | −79.8 | −73.0 | −7 | −80.0 |
| $N_2$ | | −72.0 | −70.2 | −77.5 | 7 | −70.5 |
| $N_5^+$ (term.) | −100.4[f] | | −100.4 | −103.1 | 5 | −98.1 |
| $CH_2\underline{N}N$ | | | −96.0 | −105.1 | 5 | −100.1 |
| HCN | | | −127.5 | −123.0 | −5 | −128.0 |
| $N_5H$ (N1−H) | | | | −128.5 | 7 | −121.5 |
| $N_3^-$(centr.) | −133.7 | −131[g]<br>−144 ± 3[h] | −130.4 | −131.9 | 5 | −126.9 |
| $CH_3CN$ | | | −137.1 | −132.3 | −5 | −137.3 |
| $CH_3N\underline{N}N$ | | | −129.7 | −133.9 | 5 | −128.9 |
| $HN\underline{N}N$ | | | −134.5 | −139.3 | 6 | −133.3 |
| ClCN | | | −144.0 | −145.0 | 6 | −139.0 |
| $N\underline{N}O$ | | | −140.0 | −149.9 | 6 | −143.9 |
| $N_5^+$ (N2, N4) | −165.3[f] | | −165.3 | −171.7 | 6 | −165.7 |
| $HNN\underline{N}$ | | | −179.0 | −181.6 | 6 | −175.6 |
| CH3NN$\underline{N}$ | | | −171.0 | −181.8 | 6 | −175.8 |
| $\underline{1}$2diazole | | | −173.1 | −189.9 | 17.5 | −172.4 |
| $CH_3NC$ | | | −218.0 | −235.5 | 17.5 | −218.0 |
| $\underline{N}NO$ | | | −225.0 | −244.9 | 17.5 | −227.4 |
| Pyrrole | | | −231.4 | −245.8 | 17.5 | −228.3 |
| HCNH+ | | | −235.8 | −250.5 | 17.5 | −233.0 |
| $N_5^+$ (centr) | −237.3[f] | | −237.3 | −260.8 | 17.5 | −243.3 |
| $CH_3CNH^+$ | | | −239.2 | −261.2 | 17.5 | −243.7 |
| $N_3^-$ (term.) | −281.5 | −281<br>to −282 | −280.9 | −309.1 | 17.5 | −291.6 |
| $CH_3\underline{N}NN$ | | | −321.2 | −334.9 | 17.5 | −317.4 |
| $H\underline{N}NN$ | | | −324.9 | −343.1 | 17.5 | −325.6 |
| $(CH_3)_3N$ | | | −363.1 | −370.0 | 17.5 | −352.5 |
| $CH_3NH_3^+$ | | | −361.4 | −372.0 | 17.5 | −354.5 |
| HNCO | | | −346.0 | −374.2 | 17.5 | −356.7 |
| $CH_3CH_2NH_2$ | | | −355.4 | −377.5 | 17.5 | −360.0 |
| $NH_4^+$ | −359.8<br>−361.3<br>−362.0<br>−362.7 | | −359.6 | −381.3 | 17.5 | −363.8 |
| $CH_3NCO$ | | | −365.3 | −390.9 | 17.5 | −373.4 |
| $CH_3NH_2$ | | | −377.3 | −392.8 | 17.5 | −375.3 |
| $NH_3$ | | | −380.2 | −407.2 | 17.5 | −389.7 |

[a] Reference 57. [b] Reference 45. [c] Reference 59. [d] EBBA: *N*-(*p*-ethoxybenzylidene)-*p*-*n*-butylaniline. [e] Most recent liquid phase/solution experimental value (excluding experiments in question of Schroer et al.[43] and Butler et al.[42,67] for the sake of CP correction evaluation). [f] Reference 37. [g] Reference 67, measured in clean $D_2O-CD_3OD$ solution. [h] Reference 67, measured for complexed azide anion in a solution containing $Ce^{3+}$ and $Ce^{4+}$ ions (i.e., from dearylation of 4-$MeOC_6H_4N_5$ or from $N_3^-$ anion added to a product solution after dearylation of *N*-(4-methoxyphenyl)pyrazole).

However, that does not seem yet to be possible. Consequently, crude empirical estimates are the first step.

## Conclusions

The $^{15}N$ nuclear magnetic resonance (NMR) shielding constants computed at the many body perturbation theory and the predictive coupled cluster levels are compared with the corresponding experimental values of a series of molecules to assess corrections for solution effects. Without such corrections, it is not possible to adequately interpret results of NMR shifts for molecules like $N_5^-$, where experimental observations have been questioned. The $^{15}N$ CCSD(T)/

***Table 6.*** Summary of $^{15}N$ NMR Chemical Shifts (ppm) Calculated at MBPT(2)/cc-pVDZ or MBPT(2)/cc-pVTZ Level, Scaled to Fit CCSD(T)/cc-pVTZ Chemical Shifts, and CP Corrected and Their Comparison with the Reported Experimental Values for the Molecules in Set-C

| chem shift (ppm) | Schroer[a] expt | Butler[b] expt | calcd MBPT(2)/cc-pVDZ | calcd CCSD(T)/cc-pVDZ | calcd MBPT(2)/cc-pVTZ | CCSD(T) /cc-pVTZ | CP corr factor | CP corrected chem shift |
|---|---|---|---|---|---|---|---|---|
| $MeOC_6H_4N_3$ ($N_\alpha$) | | | −304.1 | −310.1 | −290.0 | −295.9 | 17.5 | −278.4 |
| $MeOC_6H_4N_3$ ($N_\beta$) | −135.5 | | −129.0 | −166.1 | −102.4 | −139.5 | ±5 | −134.5, −144.5 |
| $MeOC_6H_4N_3$ ($N_\gamma$) | −148.2 | | −339.1 | −172.3 | −304.0 | −137.2 | ±5 | −132.2, −142.2 |
| $MeOC_6H_4N_5$ ($N_1$) | | −80 ± 2 | −132.0 | −123.0 | −96.7 | −87.8 | ±5 | −82.8, −92.8 |
| $MeOC_6H_4N_5$ ($N_2$) | | −22 ± 2<br>−28.1[c] | −105.5 | −65.4 | −67.1 | −27.0 | ±7 | −20.0, −34.0 |
| $MeOC_6H_4N_5$ ($N_3$) | | 7 ± 2 | −57.8 | −29.7 | −14.2 | 13.9 | ±7 | 20.9, 6.9 |

[a] Reference 43. [b] Reference 67. [c] Reference 68.

***Table 7.*** Comparison of $^{15}N$ NMR Chemical Shifts (ppm) Computed at CCSD(T)/pz3d2f Level with Computed Values from the Literature (Most Recent Condensed Phase Experiments Values Also Given for Comparison although the Gas Phase Experimental Data Are Not Available in the Literature)

| | MP2[a] | KS-DFT[f] | RAS-I, RAS-E[h] | B3LYP | CCSD(T) | CCSDT | experiment |
|---|---|---|---|---|---|---|---|
| $CH_3NO_2$ | 23.7[b], −71.3 | 15.3,6.5, 3.7, −0.1 | | 18.6[g] | 7.3 | | 0 |
| $N_5^-$ | −37.9 | | | −1.7[k] | −13.5 | | |
| $N_2$ | | | | | −77.5 | −79.7[m] | −70.2 |
| $CH_3CN$ | −147.3 | −99.1, −111.0, −114.7, −124.5 | | | −132.3 | | −137.1 |
| $C_4H_4NH$ (pyrrole) | −240.1 | −218.1, −220.7, −221.6, −220.3 | | | −245.8 | | −231.4 |
| NNO (centr) | | | | | −149.9 | −150.5[m] | −140.0 |
| $N_3^-$(centr) | −159.0 | | −131.9, −141.1 | | −131.9 | | −130.4 |
| NNO (term.) | | | | | −244.9 | −245.1[m] | −225.0 |
| $N_3^-$ (term.) | −323.0 | | −314.6, −313.0 | | −309.1 | | −280.9 |
| $(CH_3)_3N$ | −372.7 | −351.6, −352.9, −353.3, −353.4 | | | −370.0 | | −363.1 |
| $CH_3NH_2$ | −397.9 | −378.5, −380.2, −380.8, −380.2 | | | −392.8 | | −377.3 |
| $NH_3$ | −396.5[c], −408.1[d], −412.7 | −400.5, −402.1, −402.6, −401.6 | | −395.2[g] | −407.2 | | −380.2 |
| $MeOC_6H_4N_3$ ($N_\alpha$) | −290.0[e] | | | −310.4[k] | −295.9 | | −290.0[j], −293.0[k] |
| $MeOC_6H_4N_3$ ($N_\beta$) | −102.4[e] | | | −144.2[k] | −139.5 | | −135.5[i], −134.3[j], −137.3[k] |
| $MeOC_6H_4N_3$ ($N_\gamma$) | −304.0[e] | | | −144.3[k] | −137.2 | | −148.2[h], −146.1[j], −149.3[k] |
| $MeOC_6H_4N_5$ ($N_1$)[l] | −96.7[e] | | | −84.1[k] | −87.8 | | −80 ± 2[j] |
| $MeOC_6H_4N_5$ ($N_2$)[l] | −67.1[e] | | | −29.2[k] | −27.0 | | −22 ± 2[j], −28.1[k] |
| $MeOC_6H_4N_5$ ($N_3$)[l] | −14.2[e] | | | −18.8[k] | 13.9 | | 7 ± 2[j] |

[a] MBPT(2)/pz3d2f, this work, unless specified otherwise. [b] MP2/6-311G**, see ref . [c] See ref 63. [d] MP2/6-311G*, see ref 64. [e] MBPT(2)/cc-pVTZ. [f] In general KS-DFT uses sum over state type expression with the solutions of the coupled perturbed KS equations. The choice of exchange correlation potential is usually guided by a previous calibration step. The four values correspond to the uncoupled Kohn−Sham (KS) and three levels of refinements to uncoupled KS. See ref 65 for further details. [g] B3LYP/6-311++G**, see ref 66. [h] RAS-A and RAS-E are both MCSCF with different choices of the active spaces. [i] See ref 43. [j] See ref 67. [k] RB3LYP/6-311++G(2d,p), see ref 68. [l] See Figure 4 for numbering scheme. [m] See ref 61.

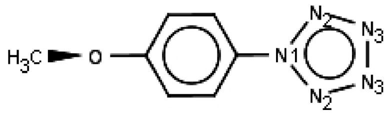


***Figure 4.*** Structure of MBPT(2)/cc-pVDZ optimized minimum for 4-methoxyphenylpentazole ($MeOC_6H_4N_5$) and its N atom labeling.

pz3d2f, pz3p2d results presented in this work expand prior benchmark CCSD(T) calculations to include $^{15}N$, and the comparison with the few known gas $^{15}N$ phase data (total of eight) establishes the accuracy of the $^{15}N$ CCSD(T) results. We observe that the CCSD(T) and MBPT(2) level shielding constant difference is small for most cases while there are a few notable exceptions. More importantly, the results are insensitive to basis set choice, and their differences can be obtained with a small cc-pVDZ basis set even when the individual values deviate significantly from the more accurate large basis set values. Since there are quite large numbers of molecules where we can obtain the MBPT(2) level chemical shifts with a suitable basis set, but not the desired CCSD(T) results, according to our findings the CCSD(T) and MBPT(2) difference can be used to correct the large basis MBPT(2) results to obtain the desired CCSD(T) estimates. Direct comparisons of the computed values with the data obtained from condensed phase experiments show a large influence from the medium and confirm that, unless the medium effects are directly incorporated in the calculations, the computed data must be corrected for the medium effects in order to be useful in practical applications. We present a systematic procedure to obtain an average correction. The error bars for the solvent effects show that for all practical purposes the same correction can be used for a range of shielding constants instead of separate corrections for each shielding constant. We show that the absolute mean error of the gas phase computed results due to the medium effects falls to 3.6 from 12.6 ppm after the proposed corrections are applied. (Within our complete set of 35 chemical shifts, the average difference between experimental liquid phase chemical shifts and the best chemical shifts calculated here was 12.6 ppm using $CH_3NO_2$ as an external standard and 8.9 ppm using $NH_3$ as an internal standard. After correcting for solvent effects, these differences decreased to 3.6 and 3.7 ppm, respectively. We are aware of the fact that, by not using $NH_3$ as an internal standard, our calculated NMR chemical shifts for N atoms in polar molecules might show slightly higher differences from experiment, and we stress this fact here. Since this study is related to the works of Schoer et al.[37,44] and Butler at al.,[42] who are using nitromethane as an external standard for all of their computed NMR chemical shifts, our choice in this study is not to use $NH_3$ as an internal standard. If the reader is interested in converting our calculated chemical shifts to those obtained by using $NH_3$ as an internal standard, the constant of 6.87 ppm should be added to all of our computed NMR chemical shifts and subtracted from all our CP correction factors.) Further work should attempt to incorporate at least pH effects into solvation models and related empirical estimates.

**Acknowledgment.** This work is supported by the United States Air Force Office of Scientific Research (AFOSR-F49620-01-1-0072).

**Supporting Information Available:** Optimized geometries of all molecules discussed in this publication. (The computation level used for each optimization is specified in the Theoretical Calculations Section.) This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Helgaker, T.; Jaszuński, M.; Ruud, K. *Chem. Rev.* **1999**, *99*, 293.

(2) Kutzelnigg, W.; Fleischer, U.; Schindler, M. In *NMR Basic Principles and Progress*; Diehl, P., Fluck, E., Günther, H., Kosfeld, R., Seelig, J., Eds.; Springer: Berlin, 1990; Vol. 23, p 165.

(3) Cremer, D.; Olsson, L.; Reichel, F.; Kraka, E. *Isr. J. Chem.* **1993**, *33*, 369.

(4) Gauss, J. *Ber. Bunsen. Ges. Phys. Chem.* **1995**, *99*, 1001.

(5) Pulay, P.; Hinton, J. F. In *Encyclopedia of Nuclear Magnetic Resonance*; Frant, D. M., Harris, R. K., Eds.; Wiley: Chichester, U.K., 1996; Vol. 7, p 4334.

(6) Facelli, J. C. *J. Phys. Org. Chem.* **2003**, *16*, 4327.

(7) Chesnut, D. B. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1996; p 245.

(8) Fleischer, U.; Kutzelnigg, W.; v. Wüllen, C. In *Encyclopedia of Computational Chemistry*; Schleyer, P.v.R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P., Schaefer, H. F., Schreiner, P. R., Eds.; Wiley: Chichester, U.K., 1998; p 1827.

(9) Bühl, M. *J. Chem. Phys.* **1993**, *99*, 1835.

(10) Gauss, J.; Stanton, J. F. *Adv. Chem. Phys.* **2002**, *123*, 355.

(11) Klemp, C.; Bruns, M.; Gauss, J.; Häussermann, U.; Stösser, G.; v.Wüllen, L.; Jansen, M.; Schnöckel, H. *J. Am. Chem. Soc.* **2001**, *123*, 9099.

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

(13) Brown, S. P.; Schaller, T.; Seelbach, U. P.; Koziol, F.; Ochsenfeld, C.; Klärner, F.-G.; Spiess, H. W. *Angew. Chem., Int. Ed.* **2001**, *40*, 717.

(14) Siehl, H.-U.; Müller, T.; Gauss, J. *J. Phys. Org. Chem.* **2003**, *16*, 577.

(15) Price, D. R.; Stanton, J. F. *Org. Lett.* **2002**, *4*, 2809.

(16) Gauss, J. *Chem. Phys. Lett.* **1992**, *191*, 614.

(17) Gauss, J. *J. Chem. Phys.* **1993**, *99*, 3629.

(18) Bartlett, R. J.; Musial, M. *Rev. Mod. Phys.* **2007**, *79*, 291.

(19) Gauss, J.; Stanton, J. F. *J. Chem. Phys.* **1995**, *102*, 251.

(20) Gauss, J.; Stanton, J. F. *J. Chem. Phys.* **1995**, *103*, 3561.

(21) Stanton, J. F.; Gauss, J.; Siehl, H.-U. *Chem. Phys. Lett.* **1996**, *262*, 183.

(22) Stanton, J. F. *J. Chem. Phys.* **1996**, *104*, 2574.

(23) Jaszuński, M.; Helgaker, T.; Ruud, K.; Bak, K. L.; Jørgensen, P. *Chem. Phys. Lett.* **1994**, *220*, 154.

(24) Barszczewicz, A.; Helgaker, T.; Jaszuński, M.; Jørgensen, P.; Ruud, K. *J. Chem. Phys.* **1994**, *101*, 6822.

(25) Malkin, V. G.; Malkina, O. L.; Salahub, D. R. *Chem. Phys. Lett.* **1994**, *221*, 91.

(26) Malkin, V. G.; Malkina, O. L.; Casida, M. E.; Salahub, D. R. *J. Am. Chem. Soc.* **1994**, *116*, 5898.

(27) Schreckenbach, G.; Ziegler, T. *J. Phys. Chem.* **1995**, *99*, 606.

(28) Rauhut, G.; Puyear, S.; Wolinski, K.; Pulay, P. *J. Phys. Chem.* **1996**, *100*, 6310.

(29) Cheeseman, J. R.; Trucks, G. W.; Keith, T. A.; Frisch, M. J. *J. Chem. Phys.* **1996**, *104*, 5497.

(30) Olsson, L.; Cremer, D. *J. Chem. Phys.* **1995**, *105*, 8995.

(31) Sundholm, D.; Gauss, J.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *243*, 264.

(32) von Philipsborn, W.; Mueller, R. *Angew. Chem., Int. Ed.* **1986**, *25*, 383.

(33) Mason, J. *Chem. Rev.* **1981**, *81*, 205.

(34) Witanowski, M.; et al. *J. Magn. Reson.* **1998**, *131*, 54.

(35) Solum, M. S.; et al. *J. Am. Chem. Soc.* **1997**, *119*, 9804.

(36) Auer, A. A.; Gauss, J.; Stanton, J. F. *J. Chem. Phys.* **2003**, *118*, 10407.

(37) Christe, K. O.; Wilson, W. W.; Sheehy, J. A.; Boatz, J. A. *Angew. Chem., Int. Ed.* **1999**, *38*, 2004.

(38) Stanton, J. F. Private communication.

(39) Ferris, K; Bartlett, R. J. *J. Am. Chem. Soc.* **1992**, *114*, 8302.

(40) Perera, S. A.; Bartlett, R. J. *Chem. Phys. Lett.* **1999**, *314*, 381.

(41) Fau, S.; Wilson, J.; Bartlett, R. J. *J. Phys. Chem. A* **2002**, *106*, 4639.

(42) Butler, R. N.; Stephens, J. C.; Burke, L. A. *Chem. Commun.* **2003**, 1016.

(43) Schroer, T.; Haiges, R.; Schneider, S.; Christe, K. O. *Chem. Commun.* **2005**, 1607.

(44) Perera, S. A.; Gregusova, A.; Bartlett, R. J. *J. Phys. Chem. A* **2009**, *113*, 3197.

(45) Levy, G. C.; Lichter, R. L. *Nitrogen-15 Nuclear Magnetic Resonance Spectroscopy*; Wiley Interscience: New York, 1979; pp 27−107.

(46) Kirkwood, J. G. *J. Chem. Phys.* **1934**, *2*, 351.

(47) Onsager, L. *J. Am. Chem. Soc.* **1936**, *58*, 1486.

(48) Udier-Blagovic, M.; De Tirado, P. M.; Pearlman, S. A.; Jorgensen, W. L. *J. Comput. Chem.* **2004**, *25*, 1322.

(49) Tzeli, D.; Mavridis, A.; Xantheas, S. S. *J. Chem. Phys.* **2000**, *112*, 6178.

(50) Aces II, University of Florida version: Stanton, J. F.; Gauss, J.; Perera, S. A.; Yau, A. D.; Watts, J. D.; Nooijen, M.; Oliphant, N.; Szalay, P. G.; Lauderdale, W. J.; Gwaltney, S. R.; Beck, S.; Balková, A.; Bernholdt, D. E.; Baeck, K.-K.; Rozycko, P.; Sekino, H.; Huber, C.; Pittner, J. Bartlett, R. J. Aces II; Quantum Theory Project, University of Florida: Gainesville, FL. Integral packages included are VMOL (Almlöf, J.; Taylor, P.R. ),VPROPS (Taylor, P. R.), and. ABACUS (Helgaker, T.; Aa. Jensen, H.J.; Jorgensen, P.; Olsen, J.; Taylor, P.R.).

(51) ACES II Mainz−Austin−Budapest version: Stanton, J. F.; Gauss, J.; Watts, J. D.; Szalay, P. G.; Bartlett, R. J.; Auer, A. A.; Bernholdt, D. B.; Christiansen, O.; Harding, M. E.; Heckert, M.; Heun, O.; Huber, C.; Jonsson, D.; Jusélius, J.; Lauderdale, W. J.; Metzroth, T.; Michauk, C.; Price, D. R.; Ruud, K.; Schiffmann, F.; Tajti, A.; Varner, M. E.; Vázquez, J. Aces II. The integral packages: MOLECULE (Almlöf, J.; Taylor, P. R.), PROPS (Taylor, P. R.), andABACUS (Helgaker, T.; Aa. Jensen, H. J.; Jørgensen, P.; Olsen, J.).

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

(53) Schäfer, A.; Horn, H.; Ahlrichs, R. *J. Chem. Phys.* **1992**, *97*, 2571.

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

(55) Witanowski, M.; Stefaniak, L.; Januszewski, H. *Nitrogen NMR*; Witanowski, M., Webb, G. A., Eds.; Plenum Press: New York, 1973; pp 245−254.

(56) Webb, G. A.; Witanowski, M. *Proc. Indian Acad. Sci. (Chem. Sci.)* **1985**, *94*, 241.

(57) Jameson, C. J.; Jameson, A. K.; Oppusunggu, D.; Wille, S.; Burrell, P. M.; Mason, J. *J. Chem. Phys.* **1981**, *74*, 81.

(58) Jameson, C. J. *Chem. Rev.* **1991**, *91*, 1375.

(59) Berger, S.; Braun, S.; Kalinowski, H.-O. *NMR-Spektroskopie von Nichtmetallen, Bd. 2*; Wiley-VCH: New York, 1992; pp 27−107.

(60) Witanowski, M.; Stefaniak, L.; Webb, G. A. In *Annual Reports on NMR Spectroscopy*; Webb, G. A., Ed.; Academic Press: London, 1977; Vol. 7, p 117.

(61) Gauss, J. *J. Chem. Phys.* **2002**, *116*, 4773.

(62) Hathaway, B. A.; Day, G.; Lewis, M.; Glaser, R. *J. Chem. Soc., Perkin Trans. 2* **1998**, 2713.

(63) Dokalik, A.; Kalchhauser, H.; Mikenda, W.; Schweng, G. *Magn. Reson. Chem.* **1999**, *37*, 895.

(64) Barfield, M.; Fagerness, P. *J. Am. Chem. Soc.* **1997**, *119*, 8699.

(65) Fadda, E.; Casida, M. E.; Salahub, D. R. *J. Phys. Chem. A* **2003**, *107*, 9924.

(66) Alkorta, I.; Elguero, J. *Struct. Chem.* **1998**, *9*, 187.

(67) Butler, R. N.; Hanniffy, J. M.; Stephens, J. C.; Burke, L. A. *J. Org. Chem.* **2008**, *73*, 1354.

(68) Burke, L. A.; Butler, R. N.; Stephens, J. C. *J. Chem. Soc., Perkin Trans.* **2001**, *2*, 1697.

CT9005739

JCTC Journal of Chemical Theory and Computation

# Vibrational Raman Spectra from the Self-Consistent Charge Density Functional Tight Binding Method via Classical Time-Correlation Functions

Steve Kaminski,*,[†] Michael Gaus,[‡] Prasad Phatak,[§] David von Stetten,[†] Marcus Elstner,[‡] and Maria Andrea Mroginski*,[†]

*Technische Universität Berlin, Institut für Chemie, Max-Volmer-Laboratorium, Sekr. PC 14, Strasse des 17. Juni 135, D-10623 Berlin, Germany, Universität Karlsruhe, Institut für Theoretische Chemie, Kaiserstrasse 12, D-76131 Karlsruhe, Germany, and Department of Chemistry, Indiana University, 800 E Kirkwood Avenue, Bloomington, Indiana 47405*

Received December 9, 2009

**Abstract:** The Self-Consistent Charge Density Functional Tight Binding (SCC-DFTB) method has been extended for the calculation of vibrational Raman spectra employing the Fourier Transform of Time-Correlation Function (FTTCF) formalism. As Witek and co-workers have already shown for a set of various organic molecules, the minimal basis SCC-DFTB approach performs surprisingly good in terms of polarizability calculations. Therefore, we were encouraged to use this electronic structure method for the purpose of Raman spectra calculations via FTTCF. The molecular polarizability was accessed via second order numeric derivatives of the SCC-DFTB energy with respect to the components of an external electric field "on-the-fly" during a molecular dynamics (MD) simulation. The finite electric field approach delivers Raman spectra that are in overall good agreement for most of 10 small organic model compounds examined in the gas phase compared to a standard Normal Mode Analysis (NMA) approach at the same (SCC-DFTB) and at a higher level of theory (BLYP aug-cc-pVTZ). With the use of reparametrized SCC-DFTB repulsive potentials, a distinct improvement of the Raman spectra from the SCC-DFTB/FTTCF protocol of conjugated hydrocarbons has been observed. Further QM/MM test calculations of L-phenylalanine in aqueous solution revealed larger deviations concerning vibrational frequencies and relative intensities for several stretching and bending modes in the benzene ring as compared to experimental results. Our SCC-DFTB/FTTCF approach was also tested against a hybrid method, in which polarizability calculations at the B3YLP 6-31G(d) level were performed on a trajectory at the SCC-DFTB level. We found that our SCC-DFTB/FTTCF protocol is not only much more efficient but in terms of the resulting Raman spectra also of similar accuracy compared to the hybrid approach. In our opinion, the more accurately calculated polarizabilities at the B3YLP 6-31G(d) level cannot compensate for the usually insufficient sampling of phase space when employing high level QM methods in a FTTCF framework.

## Introduction

Vibrational infrared and Raman spectroscopy are among the most important experimental techniques currently used to obtain information on structures and chemical states of a variety of chemical systems. In order to increase the amount of information obtained about the system under investigation, experimental spectra are often compared to those generated from electronic structure methods. A clear assignment of the observed vibrational bands to specific intramolecular motions is one of the most important contributions of computational methods to the interpretation of experimental data. Moreover, given that accurate intra- and intermolecular force fields are available, computational vibrational spectroscopy is able to give detailed insight to molecular structures and their interactions with the environment.

* Corresponding author e-mail: steve.kaminski@chem.tu-berlin.de (S.K.), andrea.mroginski@tu-berlin.de (M.A.M.).

[†] Technische Universität Berlin.

[‡] Universität Karlsruhe.

[§] Indiana University.

10.1021/ct900660x
Published on Web 02/24/2010

For such calculations, density functional methods[1,2] (DFT) are usually employed since they offer accurate results, at least for systems in the gas phase, at a reasonable computational cost. With a given electronic structure method, the vibrational spectra of molecular systems can be accessed via several rather different methodologies: the so-called Normal Mode Analysis (NMA),[3] Fourier transform of time-correlation functions (FTTCF),[4] and methods based on a principal component analysis (PCA).[5]

NMA is by far the most frequently applied method in the field of computational chemistry. According to the NMA approach, the vibrational frequencies are obtained by the diagonalization of the Hessian matrix (second derivatives of the energy with respect to the atomic displacements) at an equilibrium geometry of the molecule. The spectral intensities are independently calculated through the first spatial derivatives of either the molecular dipole moment (infrared) or the molecular polarizability (Raman). The third feature of a vibrational band, its shape, is not accessible from the NMA approach since the calculations are formally performed at 0 K.

The second methodology for calculating vibrational spectra refers to the so-called Fourier transform of time-correlation functions (FTTCF). Based on Fermi's golden rule,[4] linear response theory[4,10,11] delivers expressions for a practical calculation of infrared and Raman spectra from dipole ($\boldsymbol{\mu}$) and polarizability ($\boldsymbol{\alpha}$) time correlation functions given by

$$\text{Infrared: } I(\omega) \propto \int_{-\infty}^{\infty} \langle \boldsymbol{\mu}(t) \cdot \boldsymbol{\mu}(0) \rangle \, \mathrm{e}^{-i\omega t} \, \mathrm{d}t \tag{1}$$

$$\text{Raman: } I(\omega) \propto \int_{-\infty}^{\infty} \langle \boldsymbol{\alpha}(t) \cdot \boldsymbol{\alpha}(0) \rangle \, \mathrm{e}^{-i\omega t} \mathrm{d}\mathrm{t} \tag{2}$$

Once a time series of $\boldsymbol{\mu}$ or $\boldsymbol{\alpha}$ is available for a molecular system, the complete infrared or Raman spectrum containing frequency, intensity, and band shape information can be directly obtained from evaluating the expressions 1 and 2, respectively. These expressions can be applied under the assumption that the linear response approximation is valid; i.e., the perturbation of the system due to the applied field is small.

Besides the FTTCF formalism, another established methodology exists to extract molecular vibrations from classical trajectories. The principal component analysis, also often termed quasiharmonic analysis or essential dynamics, is based on a statistical analysis of (mass weighted) atomic fluctuations. The cross-correlator of such atomic displacements, called the covariance matrix, is usually diagonalized to yield its eigenvalues and eigenvectors, comparable to those obtained from the Hessian matrix in a standard normal-mode analysis. Thus, the method provides vibrational frequencies and normal modes of the regarded system.

The PCA technique is usually employed to identify large correlated motions in macromolecules, such as proteins,[6−8] with related modes in the far-infrared region (FIR). Wheeler and co-workers[9] were the first to evaluate the performance of this statistical method in the calculation of vibrational frequencies of small molecules in the mid infrared region (MIR). However, for this type of calculation, Schmitz and Tavan[24] found that PCA-based methods exhibit significant limitations compared to the NMA and FTTCF formalisms.

The NMA technique depends on the harmonic approximation which assumes a quadratic potential energy expression at local minima of the potential energy surface. The necessity of finding minima on the potential surface is one of the major drawbacks of the NMA approach. Especially for large flexible molecules it becomes a nontrivial task to identify all its equilibrium conformations which are sampled at finite temperature and contribute to an experimental vibrational spectrum. These problems can be avoided by employing the FTTCF method. Here, the time series of $\boldsymbol{\mu}$ or $\boldsymbol{\alpha}$ are collected from a MD trajectory at finite temperature, where most of the equilibrium conformations sampled during a sufficiently long simulation contribute to the calculated spectrum.

Despite its drawbacks, the NMA approach is more widely used in the field of computational vibrational spectroscopy than the FTTCF method. One of the reasons is that, in order to generate a single spectrum, FTTCF requires an ensemble of structures which is computationally very demanding when using high level electronic structure methods like DFT. Nevertheless, numerous studies were performed concerning the calculation of infrared and Raman spectra via FTTCF of a rather small chemical system.[12−25] For many of these studies, the Car−Parrinello[26] molecular dynamics approach was used. Another reason is related to the assignment of vibrational bands to intramolecular motions. While being a straightforward task in the framework of NMA, it became only recently available[27,28] in terms of a systematic approach for the FTTCF formalism.

To benefit from the advantages of FTTCF, it would be necessary to combine this methodology with a computationally more efficient electronic structure method. An established approximative quantum chemical method is the Self-Consistent Charge Density Functional Tight Binding (SCC-DFTB) method derived by Elstner and co-workers.[29] One of its strengths is the accurate calculation of molecular geometries comparable to higher levels of theory.[30] At the same time, the performance of SCC-DFTB comes at a considerably reduced computational cost, i.e., about 3 orders of magnitude less compared to standard DFT methods. This allows a treatment of much larger chemical systems as, e.g., biomolecules for which SCC-DFTB has already shown to deliver accurate results compared to higher level methods.[31,32]

Concerning spectroscopic properties, Witek and co-workers have already shown that vibrational frequencies[33,34] as well as Raman intensities[35] in the framework of a normal-mode analysis are satisfactorily described by SCC-DFTB.

The combination of SCC-DFTB with FTTCF has also been successfully employed for several simple chemical systems[22,36,37] in infrared studies. Furthermore, SCC-DFTB has recently been shown to deliver accurate vibrational infrared spectra of molecules in a complex protein environment.[38] Albeit these infrared studies using SCC-DFTB in combination with FTTCF are promising, this approach has to our knowledge not been used so far for the calculation of Raman spectra. In our opinion, however, the FTTCF approach could be valuable for the interpretation of various experimental

Raman studies, especially on biomolecular systems containing large floppy molecules. The size of such systems requires a combination of FTTCF with an efficient quantum chemical method such as SCC-DFTB.

Therefore, the aim of this work was the extension of the SCC-DFTB method to calculate vibrational Raman spectra via the FTTCF formalism. The implemented methodology will be tested in detail on several small organic model compounds in the gas phase. Due to the importance of an accurate prediction of condenced phase spectra, the calculated spectrum of L-phenylalanine in aqueous solution will be also compared to experimental results.

Although beyond the scope of the present work, the long-term goal is to find out whether this methodology can be helpful to interpret Raman spectra of biomolecules (especially cofactors in proteins) in a complex protein/solvent environment.

## Theoretical Approach

**The SCC-DFTB Formalism.** The SCC-DFTB approach is an approximate quantum chemical method for which an extensive description is given elsewhere in the literature.[29] Like semiempirical methods, SCC-DFTB benefits from several approximations such as avoiding the calculations of one- and two-electron integral expressions as well as taking only valence electrons explicitly into account. Therefore, a computational speedup of about 3 orders of magnitude compared to DFT is achieved.

The SCC-DFB energy based on a second-order expansion of the DFT energy with respect to density fluctuations relative to a chosen reference density is given by

$$E = \sum_{i\mu\nu} c_{i\mu} c_{i\nu} H^0_{\mu\nu} + \frac{1}{2}\sum_{A,B} \Delta q_A \Delta q_B \gamma_{AB} + E^{\text{rep}} \qquad (3)$$

where $c_{i\mu}$ are the coefficients for the minimal basis representation of confined pseudoatomic orbitals $\psi_i = \Sigma_\mu c_{i\mu}\phi_\mu$ and $H^0_{\mu\nu}$ the Hamilton matrix which depends only on the reference density. The induced charge on each atom $A$ is denoted as $\Delta q_A$; $\gamma_{AB}$ is a distance dependent function describing charge interactions, and $E^{\text{rep}}$ means a sum of two-centered core potentials. The coefficients $c_{i\mu}$ are determined by solving the Kohn−Sham equations and transforming them into a set of algebraic equations,

$$\sum_\nu c_{i\nu}(H_{\mu\nu} - \varepsilon_i S_{\mu\nu}) = 0 \qquad (4)$$

with the charge self-consistent Hamiltonian

$$H_{\mu\nu} = \langle\phi_\mu|\hat{H}^0|\phi_\nu\rangle + \frac{1}{2}S_{\mu\nu}\sum_C \Delta q_C(\gamma_{AC} + \gamma_{BC}) \qquad (5)$$

The overlap matrix elements $S_{\mu\nu}$ and the $H^0{}_{\mu\nu}$ are calculated using the PBE functional[39] and tabulated for a dense mesh of interatomic distances.

As it will be discussed in one of the following sections in more detail, we need an expression for the SCC-DFTB energy of a molecular system interacting with an external electric field to obtain a Raman spectrum via the FTTCF formalism. For this purpose, we follow Elstner[40] with the addition of an extra term to eq 3, describing the interaction of the field with the induced Mulliken charges of the system, as

$$E^{\text{field}} = E - \sum_A \Delta q_A \sum_{j=1}^{3} D_j\, x_{jA} \qquad (6)$$

Here, $D_j$ is the Cartesian component of the electric field and $x_A$ denotes the Cartesian coordinate of atom $A$. The charge self-consistent Hamiltonian (eq 5) becomes

$$H^{\text{field}}_{\mu\nu} = H_{\mu\nu} - \frac{1}{2}S_{\mu\nu}\sum_{j=1}^{3} D_j\, x_{jA} \qquad (7)$$

**The FTTCF Formalism for Raman Spectra.** The evaluation of the FTTCF formalism for Raman spectra, discussed only in brief, starts with the Raman differential scattering cross section in the quantum mechanical framework[4,10] given by

$$\lambda^4 \frac{d^2\sigma}{d\omega d\Omega} = \sum_i \sum_f \rho_i |\langle f|\varepsilon^s \mathbf{P}\cdot\varepsilon^0|i\rangle|^2 \delta(\omega_{fi} - \omega) \qquad (8)$$

where $|i\rangle$ and $|f\rangle$ denote the wave functions of the initial and final states of the system and $\rho_i$ is the probability for the system to be found in the initial state $i$. On the left side, $\lambda$ denotes the wavelength of the scattered radiation and the absorption frequency $\omega_{fi}$ is proportional to the energy levels of the final ($E_f$) and initial states ($E_i$) via $\omega_{fi} = (E_f - E_i)/\hbar$. $\mathbf{P}$ is the Cartesian polarizability tensor which can be split up into an isotropic and an anisotropic component, such as

$$\mathbf{P} = \mathbf{P}_{\text{iso}} + \mathbf{P}_{\text{aniso}} = \begin{pmatrix} \alpha_{xx} & \alpha_{xy} & \alpha_{xz} \\ \alpha_{yx} & \alpha_{yy} & \alpha_{yz} \\ \alpha_{zx} & \alpha_{zy} & \alpha_{zz} \end{pmatrix} \qquad (9)$$

For eq 8 the associated linear-response equations derived by Gordon[4,10,11] are given by

$$\left(\lambda^4 \frac{d^2\sigma}{d\omega d\Omega}\right)_{\text{iso}} = \int_{-\infty}^{\infty} \left\langle \frac{1}{3} tr[\mathbf{P}_{\text{iso}}(t)\mathbf{P}_{\text{iso}}(0)]\right\rangle e^{-i\omega t}\, dt \qquad (10)$$

$$\left(\lambda^4 \frac{d^2\sigma}{d\omega d\Omega}\right)_{\text{aniso}} = \int_{-\infty}^{\infty} \langle tr[\mathbf{P}_{\text{aniso}}(t)\mathbf{P}_{\text{aniso}}(0)]\rangle\, e^{-i\omega t}\, dt \qquad (11)$$

where $tr$ denotes the trace of a matrix. These are valuable expressions for a practical calculation of Raman spectra. Basically, only the isotropic and anisotropic polarizability tensors $\mathbf{P}_{\text{iso}}(t)$ and $\mathbf{P}_{\text{aniso}}(t)$ as a function of time are needed. The corresponding rotational invariants[11] can be defined as

$$\alpha_{\text{iso}} = \frac{1}{3}[(\alpha_{xx} + \alpha_{yy} + \alpha_{zz})] \qquad (12)$$

$$\alpha^2_{\text{aniso}} = \frac{1}{3}[(\alpha_{xx} - \alpha_{yy})^2 + (\alpha_{yy} - \alpha_{zz})^2 + (\alpha_{zz} - \alpha_{xx})^2 + (\alpha_{yy} - \alpha_{zz})^2 + 6(\alpha^2_{xy} - \alpha^2_{yz} + \alpha^2_{yz})] \qquad (13)$$

The two invariants measuring the isotropy and anisotropy of the electronic polarizability are connected to $\mathbf{P}_{\text{iso}}$ and $\mathbf{P}_{\text{aniso}}$ via

$$\frac{1}{3}tr[\mathbf{P}_{\mathrm{iso}}\mathbf{P}_{\mathrm{iso}}] = \alpha_{\mathrm{iso}} \tag{14}$$

$$tr[\mathbf{P}_{\mathrm{aniso}}\mathbf{P}_{\mathrm{aniso}}] = \alpha_{\mathrm{aniso}}^2 \tag{15}$$

The isotropic and anisotropic components of the scattering cross sections apply, in particular, to experimental spectra where the scattered light is measured parallel or perpendicular to the plane of the polarized incident laser beam[11]

$$\left(\lambda^4\frac{d^2\sigma}{d\omega d\Omega}\right)_{\parallel} = \left(\lambda^4\frac{d^2\sigma}{d\omega d\Omega}\right)_{\mathrm{iso}} + \frac{2}{15}\left(\lambda^4\frac{d^2\sigma}{d\omega d\Omega}\right)_{\mathrm{aniso}} \tag{16}$$

$$\left(\lambda^4\frac{d^2\sigma}{d\omega d\Omega}\right)_{\perp} = \frac{1}{10}\left(\lambda^4\frac{d^2\sigma}{d\omega d\Omega}\right)_{\mathrm{aniso}} \tag{17}$$

The ratio of these expressions leads to an important observable in Raman spectroscopy, the depolarization ratio $\rho$,

$$\rho = \frac{\left(\lambda^4\frac{d^2\sigma}{d\omega d\Omega}\right)_{\perp}}{\left(\lambda^4\frac{d^2\sigma}{d\omega d\Omega}\right)_{\parallel}} \tag{18}$$

To summarize, the vibrational Raman spectrum of a chemical system can be obtained by generating an ensemble of structures (indicated by $\langle ... \rangle$ in eqs 10 and 11), e.g., through molecular dynamics simulations and simultaneous electronic structure calculations at each time step to compute the polarizability tensor components (eqs 12 and 13: $\alpha_{xx}$, $\alpha_{yy}$, ...).

However, care must be taken since a single simulation usually covers only a small amount of phase space and the simulation time is usually also insufficiently small to reach an equal distribution of energy among the normal modes, leading to an erroneous intensity pattern in the spectra as pointed out, e.g., by Hornicek and co-workers.[14] More efficient than producing one rather long single simulation is the generation of multiple shorter independent trajectories, each starting with a set of randomly reassigned atomic velocities. Averaging over a certain number of spectra generally yields reliable relative intensities.

Additional errors concerning spectral intensities arise from classically (Newtonian) derived trajectories and time-correlation functions for which Berens and co-workers[11] proposed a quantum correction factor for Raman spectra,

$$Q_{hc} = \frac{\beta\hbar\omega}{1 - \exp(-\beta\hbar\omega)} \tag{19}$$

where $\beta$ is equal to $(kT)^{-1}$, $k$ denotes the Boltzmann constant, and $\omega$ is the vibrational frequency.

**Implementation of Polarizability Calculations.** In the previous section, the time dependent molecular polarizability $\mathbf{P}_{\mathrm{iso/aniso}}(t)$ was identified as the essential quantity for the evaluation of the linear response eqs 10 and 11. The components of $\mathbf{P}$ ($\alpha_{xx}$, $\alpha_{yy}$, ...) can be practically evaluated by considering the perturbation of a molecular system exposed to an external electric field $\mathbf{F}$. The energy of the perturbed system can be expressed in a Taylor series relative to the field-free energy[41] as

$$E(\mathbf{F}) = E(0) + \sum_i \left(\frac{\partial E}{\partial F_i}\right)_0 F_i + \frac{1}{2}\sum_{i,j}\left(\frac{\partial^2 E}{\partial F_i \partial F_j}\right)_0 F_i F_j ... \tag{20}$$

Assuming the harmonic approximation, the Taylor expansion is truncated after the second term which is the response property of interest, the components of the molecular polarizability,

$$\alpha_{ij} = -\left(\frac{\partial^2 E}{\partial F_i \partial F_j}\right) \qquad \text{with } i,j = x,y,z \tag{21}$$

The required second order derivatives of the energy with respect to the electric field components can be performed numerically using the following expressions[42] for diagonal

$$\alpha_{ii} = \frac{\partial^2 E}{\partial F_{ii}^2} = \frac{1}{F^2}(E_{i0} - 2E_{00} + E_{-i0}) + O(F^2) \tag{22}$$

and off-diagonal components of the polarizability tensor

$$\alpha_{ij} = \frac{\partial^2 E}{\partial F_{ij}} = \frac{-1}{2F^2}(E_{i0} + E_{-i0} + E_{0j} + E_{0-j} - 2E_{00} - E_{ij} - E_{-i-j}) + O(F^2) \tag{23}$$

where the expression for the error $O(F^2)$ can be found elsewhere.[43] The numeric step size, here equal to the electric field strength, must be chosen carefully. Too strong applied fields, on the one hand, may hamper the SCF convergence in electronic structure calculations, while a too weak perturbation may lead to numeric errors because of small energy differences. To test the numeric stability of the derived values of $\alpha_{ij}$, we follow the approach of Magdó[58] by doubling the electric field strength and evaluating $\alpha_{ij}$ again via

$$\alpha_{ij} = \frac{1}{3}\left(4\frac{\partial^2 E}{\partial F_i \partial F_j} - \frac{\partial^2 E}{\partial 2F_i \partial 2F_j}\right) \tag{24}$$

A reduced numeric error ($O(F^4)$ instead of $O(F^2)$) is obtained using eq 24, whereas the required number of energy evaluations in the presence of an electric field increases from 12 to 24.

**Vibrational Mode Assignment.** Martinez and co-workers developed a methodology[27,28] for the assignment of vibrational bands to intramolecular motions from FTTCF calculations. In the following section, we will briefly summarize the most important aspects.

The time dependent atomic velocities as available from a molecular dynamics trajectory are used as a key quantity of the methodology. The power spectra of velocity autocorrelation functions (vibrational density of states) have been used earlier to assign vibrational bands to atomic motions.[12] However, these power spectra were highly delocalized in frequency space, making an unambigious band assignment for larger molecules a difficult task. The methodology of Martinez and co-workers provides so-called effective normal modes, i.e., linear combinations of atomic displacements constructed in such a way that their corresponding power spectra are as localized as possible in frequency. To achieve this task, the following generalized eigenvalue equation must be solved:

$$\mathbf{Z}^{-1}\mathbf{K}^{(2)} = \mathbf{Z}^{-1}\mathbf{K}^{(0)}\Lambda \tag{25}$$

in which $\Lambda$ is a diagonal matrix containing the vibrational frequencies. The solution matrix $\mathbf{Z}$ performs the transformation between a set of $j$ initial velocities $\dot{x}_j$ in either Cartesian or internal coordinate space into the final set of $k$ effective normal mode velocities $\dot{q}_k$ via

$$\dot{q}_k(t) = \mathbf{Z}^{-1}\dot{x}_j(t) \tag{26}$$

Power spectra $P^q_k$ generated from these effective normal mode velocities $\dot{q}_k$ via

$$P^q_k(\omega) = \int_{-\infty}^{\infty} \langle \dot{q}_k(t)\cdot\dot{q}_k(0)\rangle\, \mathrm{e}^{-i\omega t}\, \mathrm{d}t \tag{27}$$

appear as localized bands in frequency. The matrices $\mathbf{K}^{(n)}$, necessary to set up eq 25, are calculated via

$$K_S^{(n)} = \frac{\beta}{2\pi}\int_{-\infty}^{\infty} |\omega|^n P^x_{kl}(\omega)\, \mathrm{d}\omega \tag{28}$$

Here, $P^x_{kl}$ denotes the power spectra of all auto- and cross-correlation functions from the initial set of atomic velocities in Cartesian or internal coordinate space,

$$P^x_{kl}(\omega) = \int_{-\infty}^{\infty} \langle \dot{x}_k(t)\cdot\dot{x}_l(0)\rangle\, \mathrm{e}^{-i\omega t}\, \mathrm{d}t \tag{29}$$

From the matrices $\mathbf{K}$ and $\mathbf{Z}$, an expression for the potential energy distribution (PED) can be defined,[44] i.e., the contribution of the internal coordinate $i$ to the potential energy of the $j$th normal mode,

$$\mathrm{PED}_{ij} = \sum_{k}^{3N-6} \frac{Z_{ji}Z_{ki}K^{(2)}_{jk}}{\lambda_{ii}} \tag{30}$$

## Computational Details

In order to test the performance of the implementation described in the previous section, we have chosen a set of 10 small organic molecules for *in vacuo* calculations: water ($H_2O$), butadiene ($C_4H_6$), ethanethiol ($C_2H_6S$), benzene ($C_6H_6$), methylacetate ($C_3H_6O_2$), maleimide ($C_4H_3NO_2$), *n*-methylacetamide ($C_3H_7NO$), pentane ($C_5H_{12}$), trimethylamine ($C_3H_9N$), and glycerol ($C_3O_3H_8$). In addition, the methodology was also tested for a single L-phenylalanine residue in aqueous solution.

In order to generate vibrational Raman spectra for these compounds via the FTTCF formalism, molecular dynamics simulations were performed using an extended SCC-DFTB program of Elstner and co-workers interfaced with the CHARMM[45] 32b2 software. The MD simulation protocol is described as follows for *in vacuo* simulations. In a first step, we performed an initial geometry optimization using CHARMM's adopted basis Newton−Raphson (ABNR) minimizer until a gradient threshold of 1 × $10^{-5}$ au was reached. The system was then heated and equilibrated by means of a 12.5 ps MD simulation. Subsequently, these simulations were continued for a further 400 ps under constant temperature conditions (300 K). During this production run, the fluctuating polarizability tensor elements ($\alpha_{ij}$) needed to generate Raman spectra were collected at each time step.

In the case of phenylalanine, the amino acid in its zwitterionic form (as usual at neutral pH) was placed in a cubic box of TIP3P[46] water (1485 molecules) of dimensions 35 × 35 × 35 Å$^3$. Periodic boundary conditions were applied, and electrostatic interactions beyond a cutoff radius of 12 Å were neglected by employing an atom based force-shifting function. All calculations were performed in a QM/MM framework, with phenylalanine as the QM part, as implemented in the SCC-DFTB/CHARMM interface.[47] After an initial minimization of the complete system for 1000 steps using CHARMM's ABNR minimizer, a subsequent MD simulation for the purpose of heating and equilibration was performed for 50 ps. For the following production runs (altogether 400 ps) in the NPT ensemble, the Anderson-Hoover equations[48] for constant pressure and temperature as implemented in CHARMM were employed.

Subsequent to the MD simulations, the Fourier transform of the polarizability autocorrelation was computed and further processed with a Blackman[49] filter to increase the signal-to-noise ratio. In a final step, the spectra were treated with a quantum correction factor (see eq 19) improving the intensity pattern as suggested by Berens and co-workers for Raman spectra generated from classical trajectories. Following eqs 10, 11, and 19, the Raman spectrum is finally calculated as

$$I_\omega = Q_{hc}\cdot\left[\left(\lambda^4\frac{d^2\sigma}{d\omega d\Omega}\right)_{\mathrm{iso}} + \left(\lambda^4\frac{d^2\sigma}{d\omega d\Omega}\right)_{\mathrm{aniso}}\right] \tag{31}$$

In accordance with Nyquist's theorem,[50] the spectral resolution after a Fourier transform is inversely proportional to the product of the simulation length and the chosen time step. Due to the rather low spectral density of the chosen small organic compounds for *in vacuo* calculations, a resolution of 8 cm$^{-1}$ was assumed to be sufficient. To be comparable with the experimental results, however, spectra of phenylalanine in solution were calculated at a resolution of 4 cm$^{-1}$. Hence, with a chosen time step of 0.5 fs, 8192 and 16 384 (phenylalanine) simulation steps were necessary for the generation of a single spectrum using a Fast Fourier Transform (FFT) routine. For reasons mentioned in the theory section, 100 independent MD simulations were performed per model compound, and their resulting spectra were averaged to obtain a reliable intensity pattern.

In order to evaluate the performance of the implemented SCC-DFTB/FTTCF procedure, additional Raman spectra were computed at the same level of theory (SCC-DFTB) but using the NMA technique as implemented into an SCC-DFTB standalone code by Witek et al.[33,35] In fact, the accuracy of this theoretical approach should be tested against experimental spectra of molecules in the gas phase. Unfortunately, experimental spectroscopic data recorded in the gas phase are very difficult to obtain. Thus, most of the measurements are performed in solution. Comparison to experimental data, however, will be done for the SCC-DFTB/FTTCF calculations of L-phenylalanine in water.

It should be mentioned that not only relative intensities will differ between the FTTCF and NMA approach but also vibrational frequencies, since for FTTCF they depend on the integration time step. The chosen value of 0.5 fs leads to a frequency-dependent blue shift in the spectrum according to[23]

$$\Delta\omega = \omega^3\Delta t^2/24 \tag{32}$$

The chosen step size ($\Delta t$) of 0.5 fs is believed to be a good compromise between accuracy and computational cost.

In addition, we compared the performance of SCC-DFTB/FTTCF with Raman spectra calculated at higher levels of theory. Density functional theory (DFT) provides efficient and accurate access to vibrational spectra. It has been shown[51] that reliable Raman intensities depend mostly on the size and quality of the basis set in use. Dunning's augmented triple-$\zeta$ basis set[52] (aug-cc-pVTZ) performs well in this respect. Since recent work of Riley et al.[53] also revealed very good performance in the prediction of vibrational frequencies (40 cm$^{-1}$ error) for the combination of the BLYP[54,55] functional together with Dunning's aug-cc-pVTZ basis set, we have chosen this setup in combination with an NMA approach for comparison to our SCC-DFTB/FTTCF results.

Besides our implementation, it is also possible to generate a Raman spectrum via the FTTCF formalism out of a hybrid approach. Here, we followed the procedure of Yu and Cui[22] by performing single point calculations (to obtain the molecular polarizability) at the DFT level on snapshots sampled from a trajectory computed at the SCC-DFTB level. Here, the DFT calculations were done using a computationally less demanding combination of functional and basis set (B3LYP/6-31 g(d)), since numerous single-point calculations are required (8192) to generate a spectrum. Therefore, they were restricted to only 2 of our 10 model compounds: butadiene and maleimide. With such a setup, we tried to estimate the impact of two parameters on the relative Raman intensities: namely, the size of the basis set and the number of independent trajectories used for spectral averaging. All DFT based calculations were performed using Gaussian 03.[56]

Since the output of either SCC-DFTB/NMA or DFT-BLYP/NMA calculations consists of Raman activities $A_k$ rather then intensities $I_k$, they were multiplied with a frequency dependent factor to be comparable to the SCC-DFTB/FTTCF results

$$I_k = \frac{(\omega_0 - \omega_k)^4}{\omega_k\left(1 - \exp\left[-\dfrac{h\omega_k}{K_\mathrm{b}T}\right]\right)}\cdot A_k \tag{33}$$

Here, $K_\mathrm{b}$ denotes the Boltzmann constant, $\omega_k$ is the vibrational frequency of mode $k$, and the incident laser line $\omega_0$ was chosen at a value of 1064 nm. The temperature has been chosen to be 300 K.

An important aspect concerning the interpretation of the calculated spectra is the assignment of Raman active bands to intramolecular motions. In the framework of NMA, the eigenvectors of the Hessian matrix, referring to vibrational motions, are calculated by default. Since the *in vacuo* SCC-DFTB/FTTCF calculations will be compared to calculations using the NMA approach, the eigenvectors will be used for a qualitative assignment of bands to internal motions.

The solution spectrum of phenylalanine from SCC-DFTB/FTTCF calculations has been compared to experimental data. Therefore, effective normal modes as described previously were evaluated, and the potential energy distribution of selected modes were estimated via eq 30. The atomic displacements of phenylalanine were expressed in the internal coordinate space. For this purpose, a set of 63 nonredundant internal coordinates for the amino acid were defined from bond length, bond angles, and proper and improper (out-of-plane) torsion angles following the rules of Pulay et al.[57]

## Experimental Procedure

The vibrational Raman spectrum of L-phenylalanine in water (~0.18 M) was measured at two temperatures (298 and 133 K) and pH ~ 7 using a Bruker RFS 100/S Fourier-transform spectrometer. The excitation line was at 1064 nm, and the spectrum has been recorded at a resolution of 4 cm$^{-1}$.

## Results and Discussion

**Numeric Stability of Calculated Polarizabilities.** For all 10 model compounds, MD simulations were performed to find appropriate numeric step sizes (the applied field strength) for the computation of the polarizability tensor components. Magdó et al.[58] suggested values of about 0.004 au for polarizability calculations on linear tetrapyrroles at local minima and in the framework of a NMA approach. According to the FTTCF methodology, however, the computation of polarizabilities are performed on structures extracted out of a MD trajectory, which are usually not in a local minimum. Therefore, we tested the step sizes 0.0005, 0.001, 0.005, 0.01 and 0.05 au. For all model compounds except glycerol, methylacetate, and maleimide, the effects were negligible, meaning that the spectra could not be visually distinguished from each other. For the mentioned compounds, however, the smallest step size (0.001) leads to slight artificial baseline drifts, while the spectra resulting from the two higher step sizes are almost identical. Raman spectra of methylacetate and glycerol, estimated using different step sizes, are shown in Figure 1. A pronounced baseline effect together with a decreasing signal-to-noise ratio is clearly visible for the spectrum obtained using a step size of 0.0005 au. Large step sizes (0.05 au) on the contrary only have quite small influences on the overall spectral shape.

Whereas these calculations have been performed using a simple numeric differentiation scheme (eq 21), additional tests for all model compounds were done using a more accurate approach as shown in eq 24. A comparison between the two methodologies revealed a visually negligible difference for the resulting spectra. Therefore, to keep the computational effort low, all simulations in this work have been performed using eq 21.

**Stability of Spectral Intensities.** As pointed out in the theory section, only an average over spectra from several

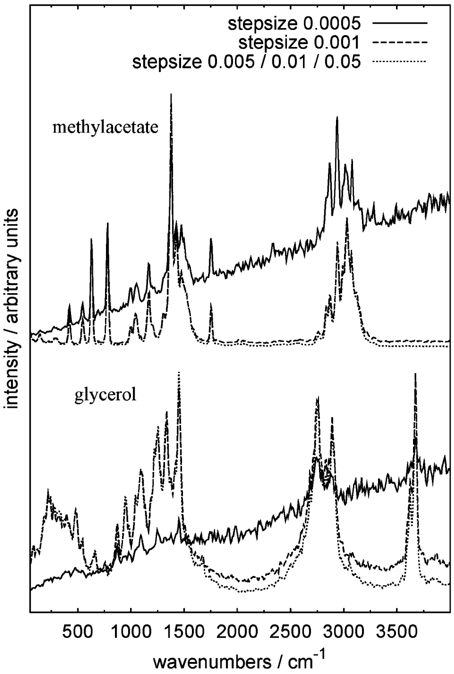


***Figure 1.*** Raman spectra of methylacetate and glycerol in dependence of numeric step sizes for the evaluation of the polarizability tensor elements. The resulting Raman spectra for step sizes of 0.005, 0.01, and 0.05 au can hardly be distinguished and are therefore represented by a single dotted line.

independent MD simulations will give reliable relative band intensities. Figure 2 exemplifies for two model compounds (pentane, *n*-methylacetamide) that spectra generated from single trajectories (bottom spectra) are by no means representative. The top spectra, on the other hand, show that in all of the spectral regions, changes in intensities between 50 and 100 on the average remain very small, so that convergence in the intensity pattern has been achieved.

**Raman Spectra in the Gas Phase.** Among the 10 model compounds for which *in vacuo* calculations have been performed, five of them have been selected for a more detailed discussion. For the remaining model compounds, only the spectra will be shown. All spectra derived from the time-correlation-function formalism as implemented in this work (termed SCC-DFTB/FTTCF) result from simulations with a step size of 0.005 au and from an average of 100 single spectra.

For the purpose of a better visual comparison with the SCC-DFTB/FTTCF results, the line spectra from the NMA approach have been convoluted using Lorentzian functions with a half-width of 10 cm$^{-1}$ and a peak maximum at the position of the calculated vibrational frequencies. In all vibrational spectra shown, the most intense band was scaled to unity.

**Water.** For the smallest model compound, the resulting spectra in Figure 3 (top left) show excellent agreement when comparing the NMA and FTTCF formalisms for the SCC-DFTB method in terms of vibrational frequencies. Compared to BLYP aug-cc-pVTZ, the highest frequency mode, i.e., the asymmetric O−H stretching, is substantially blue-shifted (244 cm$^{-1}$), whereas the other two modes are in good agreement with SCC-DFTB results. Concerning the relative intensities, very good agreement is found between SCC-DFTB/FTTCF and BLYP aug-cc-pVTZ, wheras SCC-DFTB/NMA shows a distinctly different pattern.

**Glycerol.** Analogous to water, the calculated vibrational frequencies compare well (see Figure 4 /bottom right) between SCC-DFTB/FTTCF and SCC-DFTB/NMA. This is also true for most spectral regions when incorporating BLYP aug-cc-pVTZ into the comparison. The largest spectral shift (~160 cm$^{-1}$) can be found for the double band feature in the C−H stretching region near 3000 cm$^{-1}$. A striking feature of the glycerol spectrum from the SCC-DFTB/FTTCF approach is its broad background. It is the only model compound where intramolecular hydrogen bonds appear with a continuous variation of donor−acceptor distances (O−H···O) during the MD simulations, which is most likely responsible for the observed broadening. In terms of the spectral shape, the SCC-DFTB/NMA approach is in closer agreement to BLYP aug-cc-pVTZ in the high frequency region above 3500 cm$^{-1}$. Here, the O−H stretching vibrations are overestimated in the SCC-DFTB/FTTCF spectrum. Below 500 cm$^{-1}$, O−H wagging modes at 249 and 262 cm$^{-1}$ are overestimated for the SCC-DFTB/NMA approach.

**Ethanethiol.** For the only sulfur compound, good agreement in band positions is found for all three compared methodologies in the region between 500−1500 cm$^{-1}$, as shown in Figure 4. The largest deviation comparing SCC-DFTB and DFT occurs for the S−H stretching mode at 2538 cm$^{-1}$ (SCCDFTB/NMA) which is blue-shifted by 130 cm$^{-1}$. The intensity pattern, however, varies strongly between SCC-DFTB/FTTCF and SCC-DFTB/NMA for three distinct modes. These are the ones at 139, 684, and 3065 cm$^{-1}$ (referring mainly to S−H wagging, S−C stretching, and C−H stretching vibrations), which are over- (139 cm$^{-1}$) and underestimated (684 and 3065 cm$^{-1}$), respectively, for the SCC-DFTB/NMA approach compared to the other two methodologies. In the high frequency region above 2500 cm$^{-1}$, the spectral shapes of both SCC-DFTB approaches are not in accordance with BLYP aug-cc-pVTZ.

**Maleimide.** While SCC-DFTB/FTTCF shows good agreement concerning band positions (Figure 3) over the whole spectral range compared to SCC-DFTB/NMA and BLYP aug-cc-pVTZ, its spectral shape is clearly distinct from BLYP aug-cc-pVTZ, especially the N−H stretching mode at ~3500 cm$^{-1}$, just as the region below 1200 cm$^{-1}$ is overestimated with respect to the most pronounced band at 1771 cm$^{-1}$. On the contrary, the SCC-DFTB/NMA compares well concerning relative intensities to BLYP aug-cc-pVTZ over the entire spectral range.

***n*-Methylacetamide.** Both SCC-DFTB approaches deliver a very similar vibrational spectrum concerning band positions and the overall spectral shape, as illustrated in Figure 5. The agreement with BLYP aug-cc-pVTZ in terms of band positions is also very good, except in the C−H stretching region around 3000 cm$^{-1}$. Overestimated intensities for both SCC-DFTB methods can be found in the broad spectral feature near 1450 cm$^{-1}$ mainly belonging to C−H methly deformations. In the C−H stretching region, the bands from

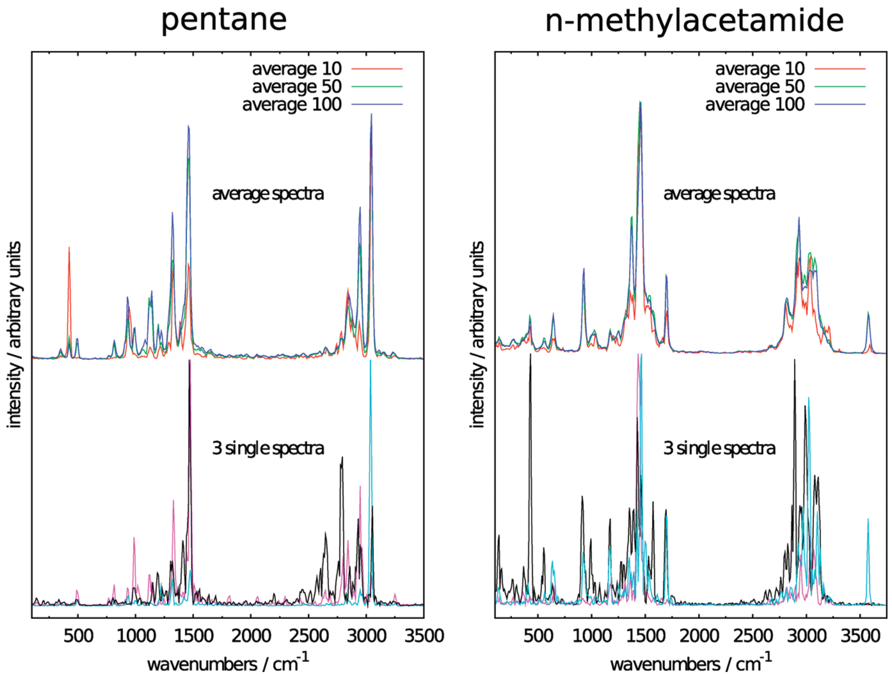


***Figure 2.*** Raman spectra from the SCC-DFTB/FTTCF formalism of two model compounds. The top spectra illustrate the convergence of the overall spectral shape as the number of single spectra for averaging increases. On the bottom, three spectra from single independent trajectories are shown to illustrate spectral variations.

both SCC-DFTB approaches are underestimated compared to BLYP aug-cc-pVTZ.

**Raman Spectra in Solution.** In Figure 6, experimental and calculated Raman spectra of L-phenylalanine in water are shown. Since the experimental spectrum recorded at room temperature, which is principally the correct one to compare to our calculations, exhibits an intense background below 750 cm$^{-1}$, it has been measured again at ~133 K. The observed background is most likely due to couplings of the solutes' vibrations to the librational motions of water which are largely suppressed at 133 K. As a result, in the low temperature spectrum, a better resolution of the solutes' vibrations in this region is observed. Therefore, we will refer to this 133 K spectrum in the following discussion.

Concerning intensity fluctuations arising from the MD simulation, small variations are found between spectra generated out of 50 and 100 independent trajectories. This is basically the same observation as for the *in vacuo* calculations, indicating a sufficient sampling of the solute in aqueous solution.

The overall spectral pattern obtained from SCC-DFTB/FTTCF calculations on L-phenylalanine fits the experiment to an extent that makes it possible to qualitatively assign most of the spectral regions, as illustrated by dashed vertical lines in Figure 6. Especially the experimental line shapes are mostly well reproduced. However, several regions in the calculated spectrum deviate significantly in terms of vibrational frequencies and/or relative intensities as compared to the experiment. These regions are labeled 1−5 in Figure 6 and will be discussed in more detail. For the purpose of an assignment of the labeled bands to vibrational motions, effective normal modes have been calculated for selected bands as shown in Figure 7. Here, the colored spectra denote the normal modes localized in frequency and therefore are helpful for an assignment of Raman active vibrational bands.

In spectral region 1 in Figure 6, the bands are overestimated by the SCC-DFTB/FTTCF calculations compared to the experiment and blue-shifted by ~80 cm$^{-1}$. From a PED analysis (eq 30), these bands are related to C−H out-of-plane motions of the benzene ring as well as C−C backbone stretchings and C−C−C bendings of the benzene ring.

The intensity of the single band marked as number 2 in Figure 6, which can be assigned to the most prominent one in the experiment, is significantly underestimated by the SCC-DFTB/FTTCF calculations, and its vibrational frequency is blue-shifted by ~55 cm$^{-1}$. The related motions are mainly C−C−C bendings of the benzene ring.

The relative intensities between the three Raman active bands in region 3 are well reproduced by the SCC-DFTB/FTTCF calculations. The intensity of this spectral region compared to the neighboring ones, however, is overestimated and blue-shifted by ~90 cm$^{-1}$. The most prominent band in the SCC-DFTB/FTTCF spectrum can be assigned to C−C stretchings and C−C−H bending motions in the benzene ring. The second and third bands are mainly composed of C−C−H bending motions in the benzene ring as well as backbone C−C and C−N stretchings.

The spectral region number 4 is characterized by a blue-shifted (~60 cm$^{-1}$) broad feature in the SCC-DFTB/FTTCF calculations. Its intensity is overestimated, and the related motions are C−C−H backbone bending and C−C as well as C−O backbone stretching motions.

The highest frequency modes in the spectrum in Figure 6 corresponding to region 5 are extremely shifted to higher wavenumbers (~200 cm$^{-1}$) as compared to the experiment. This double band feature is dominated by C−C stretching motions in the benzene ring. This is not surprising since it

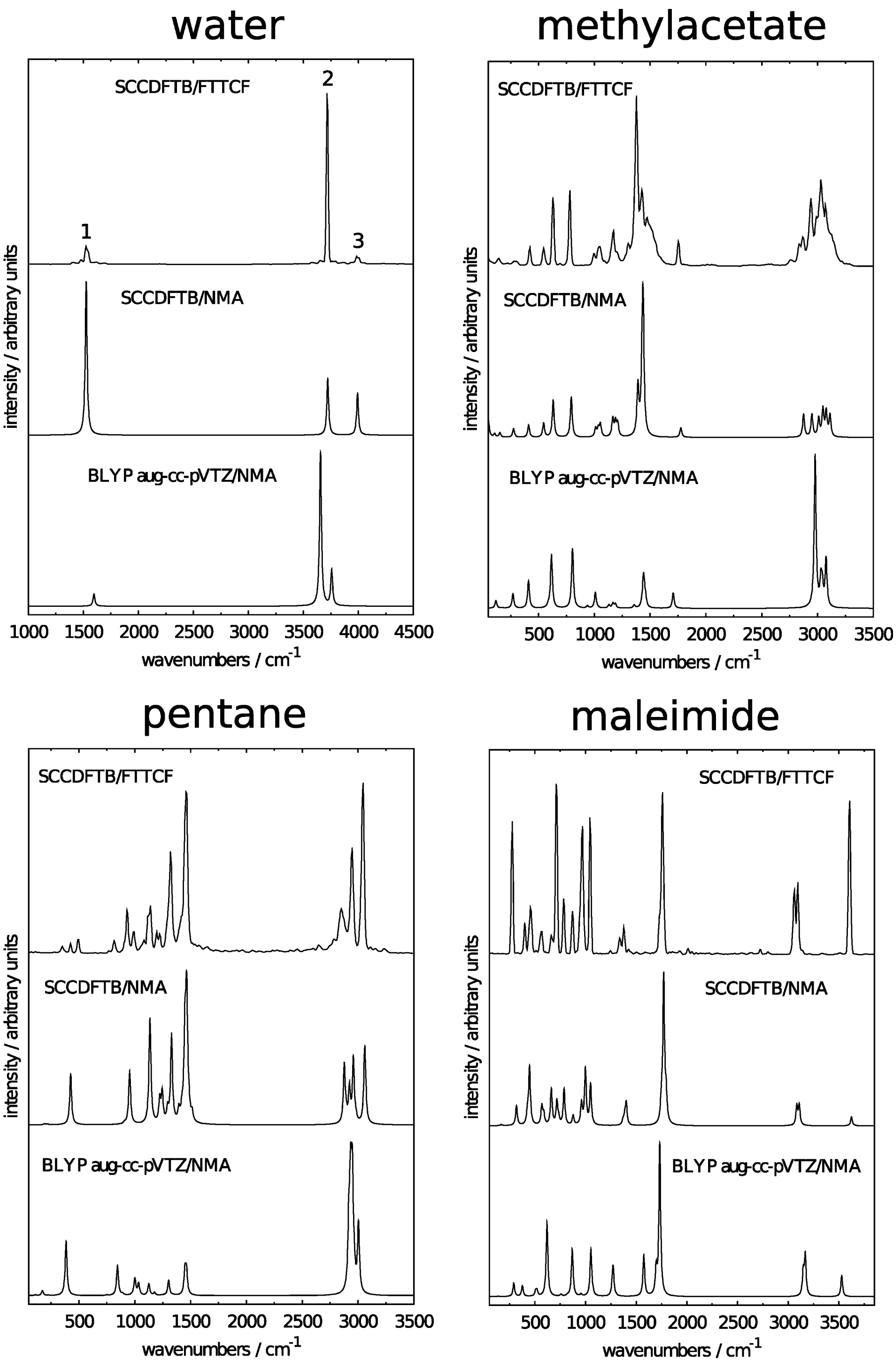


***Figure 3.*** Vibrational Raman spectra of 4 model compounds from different methodologies (FTTCF vs NMA) and different levels of theory (SCC-DFTB vs DFT) for comparison. Raman active modes of water for which depolarization ratios were estimated are marked.

is well-known for the non-self-consistent DFTB method that for benzene in the gas phase the C−C stretching mode with symmetry $E_{2g}$ is overestimated by more than 200 $\mathrm{cm}^{-1}$.[59] For the self-consistent scheme, this shortcoming is not eliminated, as shown for benzene (mode 5 in Figure 4). For SCC-DFTB, this band appears at 1826 $\mathrm{cm}^{-1}$, whereas for BLYP it is found at 1571 $\mathrm{cm}^{-1}$.

**SCC-DFTB Repulsive Potentials for Vibational Spectra Calculations.** The vibrational Raman spectra from SCC-DFTB/FTTCF calculations presented so far in this work have shown to be in overall good agreement with higher level theoretical methods and experimental data. However, vibrational bands referring to C−C stretching and bending motions in conjugated $\pi$ systems (butadiene, benzene, phenylalanine) show large frequency shifts up to 200 $\mathrm{cm}^{-1}$. Overpolarization effects in conjugated systems, a known problem of SCC-DFTB,[60] are responsable for such errors.

To overcome such problems, the improvement of SCC-DFTB repulsive pair potentials for a better prediction of a variety of molecular properties is in progress. Małolepsza and co-workers[61] developed a set of pair potentials which substantially improve calculated vibrational frequencies for modes where hydrogen and carbon atoms are involved. Furthermore, Gaus and co-workers[62] recently presented a procedure for an automatized parametrization of repulsive potentials for several molecular properties.

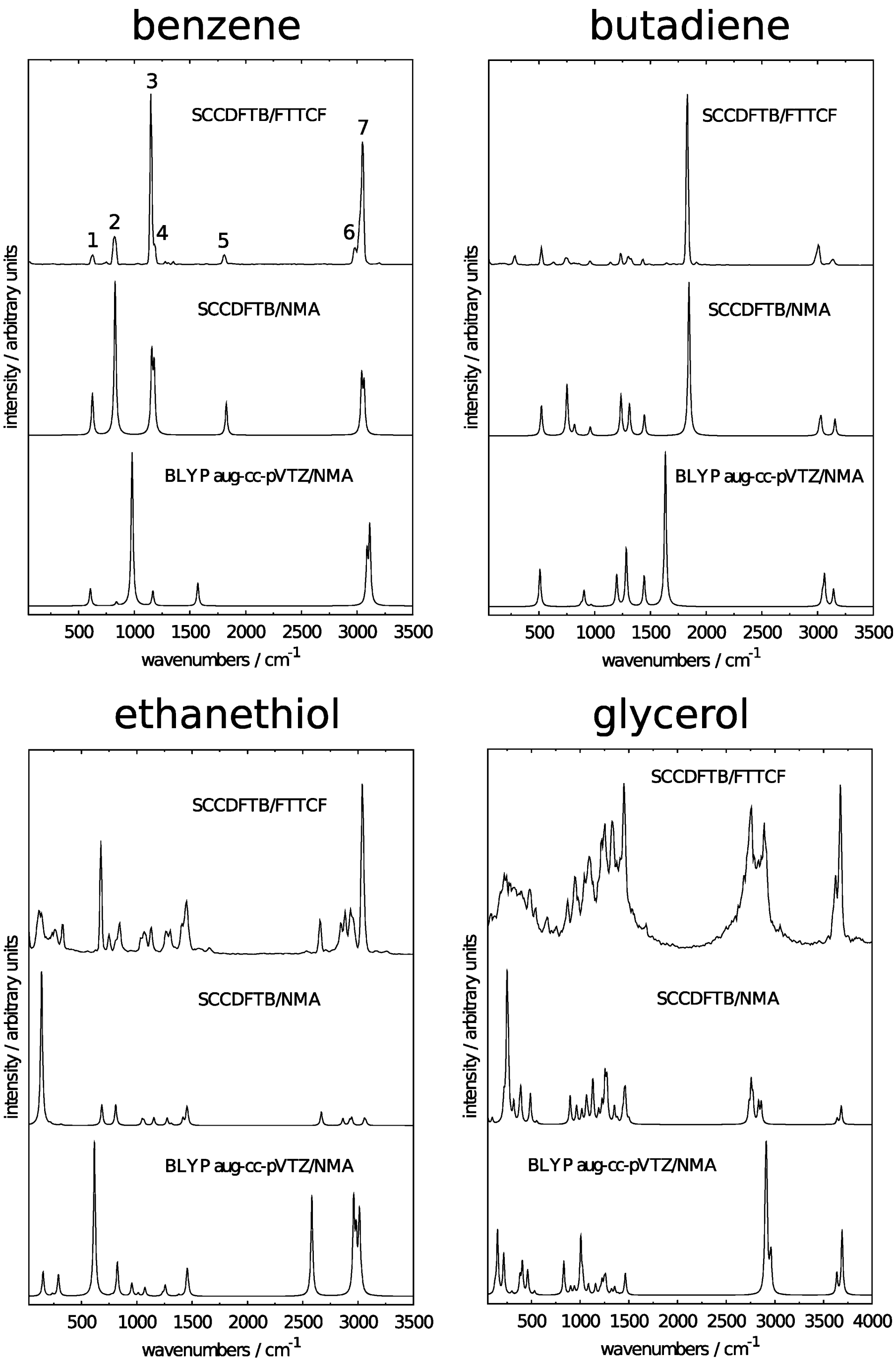


***Figure 4.*** Raman spectra of 4 model compounds from different methodologies (FTTCF vs NMA) and different levels of theory (SCC-DFTB vs DFT) for comparison. Raman active modes of benzene for which depolarization ratios were estimated are marked.

To investigate the effect of parametrized pair potentials on vibrational frequencies, we recalculated Raman spectra of benzene and butadiene (since optimized parameters only exist for carbon and hydrogen so far) in the gas phase with the SCC-DFTB/FTTCF protocol. The results (termed new-sk (Slater−Koster) parameters) shown in Figure 8 have been compared to SCC-DFTB/FTTCF calculations with the standard set of repulsive potentials (top spectra). Vibrational bands significantly affected by the new parameters in use were marked with symbols (triangle, star, circle). For benzene, three strongly shifted bands can be observed as shown in Figure 8. The resulting spectrum is in much closer agreement to the one from high level density functional methods (first two spectra from the bottom). This is also true for butadiene. Here, the intense C=C stretching mode (marked with a triangle) is red-shifted by approximately 200 cm$^{-1}$ and compares well to the results obtained from DFT calculations.

The Raman spectra from SCC-DFTB/FTTCF calculations have been compared to a variety of DFT methods with different basis sets and density functionals (first two spectra from the bottom) involved. With the new repulsive pair potentials for carbon and hydrogen atoms, most of the observed Raman active bands obtained from the SCC-DFTB/FTTCF calculations are in terms of frequency positions and intensities within the range of scattering observed from different high level DFT methods.

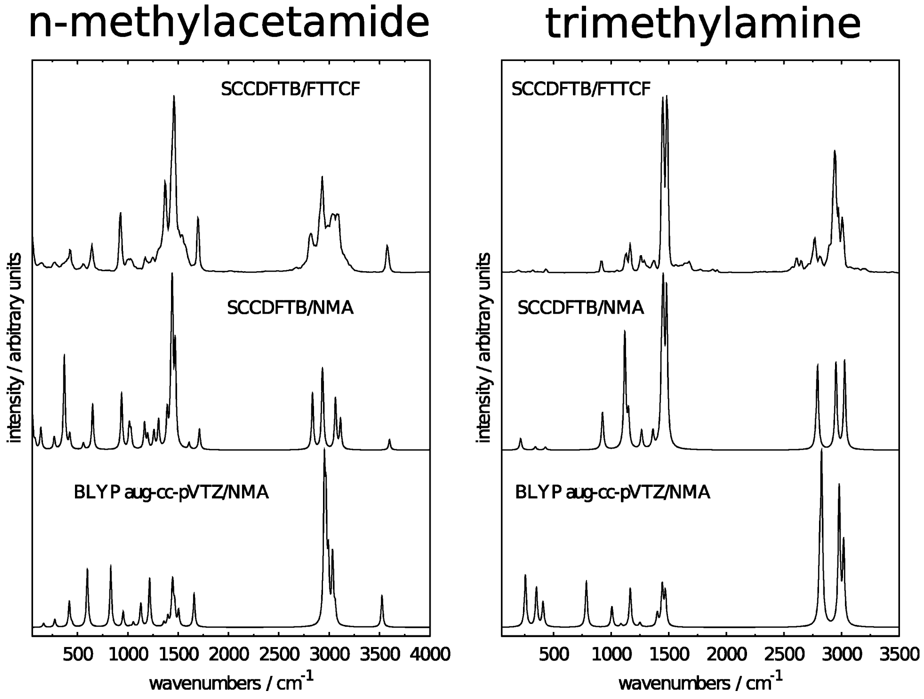


***Figure 5.*** Raman spectra of 2 model compounds from different methodologies (FTTCF vs NMA) and different levels of theory (SCC-DFTB vs DFT) for comparison.

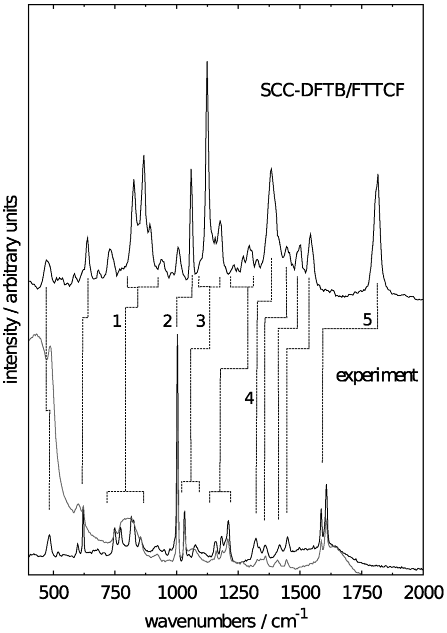


***Figure 6.*** Raman spectra of phenylalanine in water from SCC-DFTB/FTTCF calculations (top) and experimental measurements (bottom) at 298 K (light gray) and 133 K (black), respectively. The top spectrum results from an average of 100 independent MD simulations. Dashed lines indicate the qualitative assignment of various spectral regions of which the numbered ones will be discussed in the text.

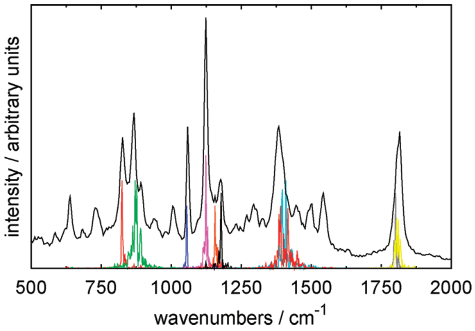


***Figure 7.*** Calculated vibrational Raman spectrum of phenylalanine in solution from SCC-DFTB/FTTCF calculations averaged over 100 independent simulations (black). Colored spectra denote localized effective normal modes referring to spectral regions numbered in Figure 6. Relative intensities of effective normal modes were manually adjusted for a better illustration.

Further improvements concerning vibrational spectra of molecules containing various functional groups can be expected, since the work on repulsive potentials for other pairs of elements is in progress.

**Depolarization Ratios.** Depolarization ratios are important observables in Raman spectroscopy. By employing eqs 16 and 17, depolarization ratios for individual vibrational modes are accessible via the FTTCF formalism and shall be compared here to results obtained from SCC-DFTB and BLYP aug-cc-pVTZ calculations for polarized laser light following the NMA approach. The respective calculations have been performed for the most prominent Raman active modes (marked in Figures 3 and 4) of water and benzene since they show the most simple spectral pattern of all tested compounds, making an unambiguous mode assignment straightforward. Due to the high symmetry of benzene ($D_{6h}$), several of its Raman active modes are degenerated. Since they are equal concerning their vibrational frequencies and depolarization ratios, each couple of degenerated modes was treated as a single band and not counted twice.

In the case of water and for the symmetric and asymmetric O−H stretching vibrations (modes 2 and 3), the depolarization ratios from SCC-DFTB/FTTCF are in good agreement with the results obtained from BLYP aug-cc-pVTZ calcula-

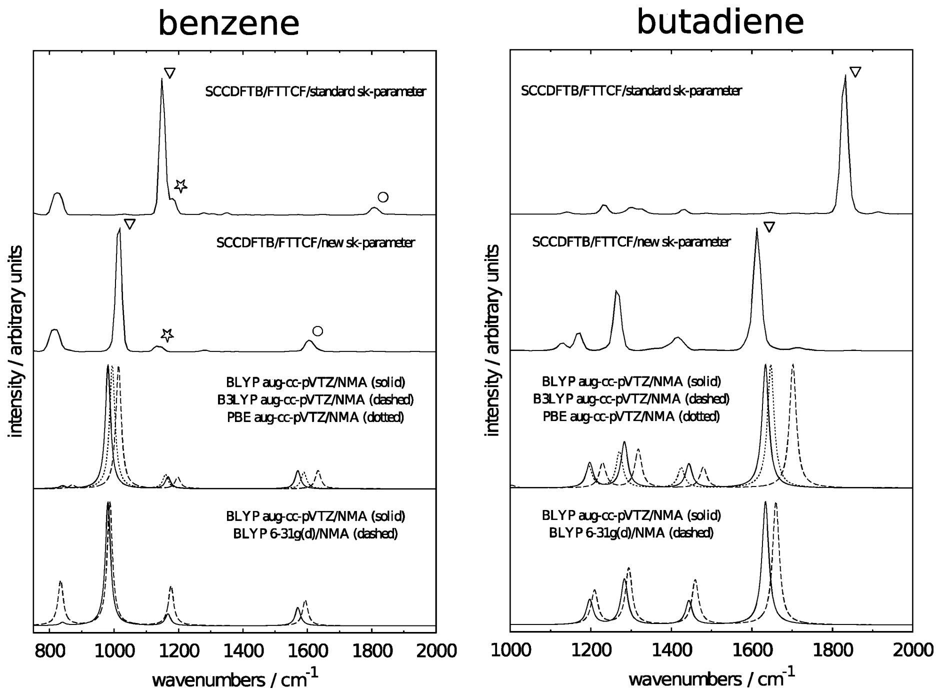


*Figure 8.* Calculated Raman spectra of butadiene and benzene in the gas phase. The first two spectra from the top of the graph result from SCC-DFTB/FTTCF calculations employing different Slater−Koster parameters for C−C and C−H repulsion. The symbols (triangle, star, circle) illustrate the frequency shift of several vibrational bands. The two graphs from the bottom show Raman spectra from DFT/NMA calculations done with different combinations of density functionals and basis sets.

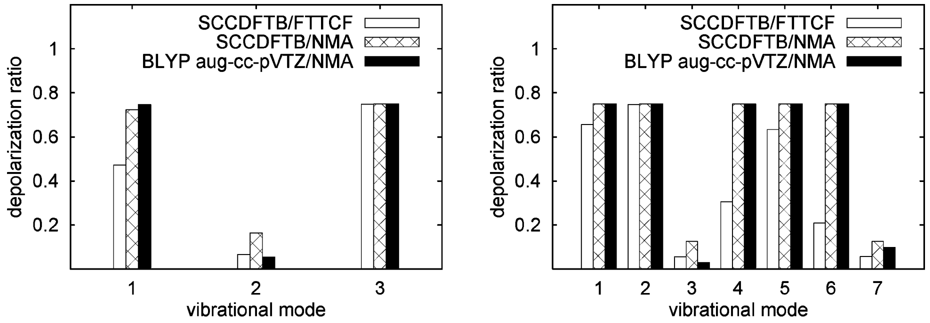


*Figure 9.* Calculated depolarization ratios for the most prominent Raman active modes (marked in Figures 3 and 4) of water (left) and benzene (right) from three different approaches as indicated in the graphs.

tions as illustrated in Figure 9. Concerning the H−O−H bending vibration (mode 1), the respective ratio from SCC-DFTB/FTTCF calculations is underestimated by approximately 25% in comparison to BLYP aug-cc-pVTZ, whereas SCC-DFTB/NMA perfectly agrees with BLYP aug-cc-pVTZ for modes 1 and 3, while mode 2 is overestimated by a factor of 3.

For benzene, depolarization ratios for all vibrational modes except number 3 are in a very good agreement comparing SCC-DFTB/NMA with BLYP aug-cc-pVTZ. The skeletal breathing mode number 3 is overestimated by the SCC-DFTB/NMA method. The SCC-DFTB/FTTCF calculations on the other hand reveal significantly reduced ratios for modes 4 and 6, referring to C−H wagging and C−H stretching vibrations, respectively.

**Raman Spectra via DFT Polarizability Calculations.** For all previously presented Raman spectra in the framework of our SCC-DFTB/FTTCF method, both the trajectory as well as the molecular polarizabilities were calculated at the SCC-DFTB level of theory. While SCC-DFTB is known to yield molecular structures in close agreement to higher level methods,[30] calculated polarizabilities are less accurate compared to high level methods due to the minimal basis set employed by SCC-DFTB. Therefore, as an alternative to our SCC-DFTB/FTTCF protocol, we figured out the performance of an approach, called the hybrid method, in which the trajectory and the polarizabilities were calculated on different levels of theory. Whereas the trajectory was still generated at the SCC-DFTB level, subsequent single-point calculations on snapshots of the trajectory were performed using higher level methods to obtain the molecular polarizabilitiy. We used such a hybrid approach to estimate the impact of the two parameters, i.e., the quality and size of the basis set for polarizability calculations on the one hand and the amount of phase space sampling on the other hand, on the overall spectral pattern.

Within our SCC-DFTB/FTTCF protocol using “on-the-fly” calculations of molecular polarizabilities, it is a computationally feasible task to generate a large ensemble of trajectories to guarantee for a sufficient phase pace sampling,

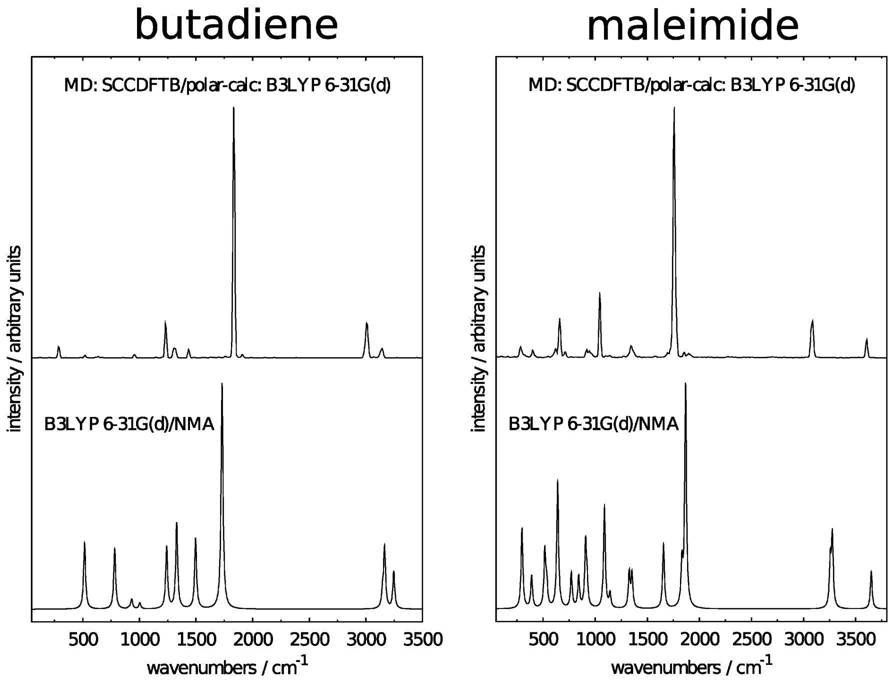


***Figure 10.*** Comparison of vibrational Raman spectra of butadiene and maleimide (each from two different approaches as explained in the text). The top spectra result from an average of spectra from 10 independent simulations.

important for a reliable spectral pattern as already shown in Figure 2. This is not true any longer when employing density functional methods, such as DFT/B3LYP 6-31G(d), for the calculation of polarizabilities. For a butadiene molecule in the gas phase, the computational effort on a conventional desktop PC to generate a Raman spectrum out of a single trajectory (8192 simulation steps/0.5 fs time step/8 $cm^{-1}$ spectral resolution) is as follows. Within the use of our SCC-DFTB/FTTCF protocol, the concurrent generation of the trajectory and the polarizabilities takes approximately 3 min on a single processing core. For the polarizability calculations itself, about 135 h of computing time is needed when employing DFT/B3LYP 6-31G(d) on a single CPU.

Due to the substantially increased computational cost, DFT/B3LYP 6-31G(d) single-point calculations were necessarily done on snapshots from a smaller ensemble of SCC-DFTB trajectories. Vibrational Raman spectra averaged over 10 single spectra were generated for two chosen model compounds. We used butadiene and maleimide as test cases, for which the resulting spectra are shown in Figure 10 and compared to a related B3LYP 6-31G(d)/NMA spectrum. Further on, we will denote the spectra derived from B3LYP 6-31G(d) single-point calculations of SCC-DFTB/FTTCF snapshots as B3LYP 6-31G(d)/polar.

Concerning butadiene, in the region above 1000 $cm^{-1}$, the B3LYP 6-31G(d)/polar calculations deliver a spectrum in good agreement to its B3LYP 6-31G(d)/NMA analogue, as shown in the left picture of Figure 10. The respective bands mainly refer to C−C and C−H stretching vibrations. However, the spectral pattern below 1000 $cm^{-1}$ is not reproduced satisfactorily. Here, from the B3LYP 6-31G(d)/polar protocol, the intensities of the vibrational bands are strongly underestimated.

The observations made for butadiene are also valid for maleimide. The most pronounced C=O stretching vibration at 1833 $cm^{-1}$ (B3LYP 6-31G(d)/NMA) as well as the C−H and N−H stretchings in the region above 3000 $cm^{-1}$ are in a very good agreement compared to the B3YLP 6-31G(d)/NMA spectrum, whereas the SCC-DFTB/FTTCF calculations (Figure 3) show a very different spectral pattern in this region. The relative intensities below 1500 $cm^{-1}$ are again not satisfactory within the B3LYP 6-31G(d)/polar approach compared to the results from NMA calculations, indicating that a sufficient sampling has not been achieved after an overall simulation time of 40 ps.

The fact that a 40 ps MD simulation is not necessarily sufficient in terms of the Raman intensity pattern is also illustrated for pentane in Figure 2. The spectrum colored in red (40 ps) differs in the region below 1500 $cm^{-1}$ significantly from the ones in green (200 ps) and blue (400 ps). Nonella and co-workers[25] made a similar observation by comparing infrared Spectra from an FTTCF and NMA approach of *p*-benzoquinone in aqueous solution. They concluded that the 17.5 ps QM/MM MD simulation, although consuming considerable computational resources, was too short for the computation of a reliable vibrational spectrum.

Larger spectral deviations in the lower frequency region can also be observed by the comparison of the B3LYP 6-31G(d)/NMA spectra to their analogues (BLYP aug-cc-pVTZ) in Figures 3 and 4. The spectral pattern in this region is therefore highly sensitive to the level of theory on the one hand, as well as, in the FTTCF framework, to the number of independent trajectories used for spectral averaging on the other hand.

From our test calculations, we conclude that a hybrid approach to calculate Raman spectra via the FTTCF formalism yields results which are not clearly superior to the ones from our SCC-DFTB/FTTCF protocol, at least not below 1500 $cm^{-1}$. The generation of spectra via the hybrid approach is, however, computationally much more demanding and therefore, in contrast to the SCC-DFTB/FTTCF method, restricted to rather small chemical system.

## Summary and Conclusions

The SCC-DFTB electronic structure method has been extended for "on-the-fly" calculations of molecular polarizabilities via a finite electric field approach to access vibrational Raman spectra in the framework of the FTTCF formalism. FTTCF, in contrast to a standard Normal Mode Analysis, incorporates anharmonic motions as well as effects from the fluctuating environment at a finite temperature in the pattern of a vibrational spectrum.

The numeric differentiation step size for the calculation of polarizabilitiy tensor elements has been shown to be important, and a value of 0.005 au or larger was sufficient to avoid obvious numeric errors for the set of tested molecules. Furthermore, the stability of the employed numeric differentiation scheme has been verified to be sufficient by comparison to a more sophisticated method proposed by Magdó and co-workers.

Vibrational Raman spectra generated via the SCC-DFTB/FTTCF formalism were in good agreement for a set of 10 model compounds examined in the gas phase and compared to an NMA approach at the same (SCC-DFTB) and at a higher level of theory (BLYP aug-cc-pVTZ). Especially the consensus of vibrational frequencies for both methodologies (NMA vs FTTCF) at the SCC-DFTB level suggests that an integration time step of 0.5 fs is not too coarse and is a good compromise between accuracy and computational effort.

Compared to high level BLYP aug-cc-pVTZ calculations, the largest deviations concerning vibrational frequencies using SCC-DFTB were found for C−C stretchings in conjugated systems as well as C−H stretching modes in general. We were able to correct for such errors by employing a set of optimized repulsive pair potentials for SCC-DFTB. Recalculated Raman spectra for butadiene and benzene with the SCC-DFTB/FTTCF protocol were in very good agreement compared to high level desity functional methods. With ongoing progress in parametrizing SCC-DFTB repulsive potentials for other pairs of elements, further improvements in vibrational spectra calculations can be expected.

With respect to the relative band intensities compared to spectra from high level methods, none of the two techniques (neither NMA nor FTTCF) could be identified to be clearly superior to the other one in combination with the SCC-DFTB method. The situation varies rather strongly in terms of the specific compound and the spectral region under consideration. Concerning the FTTCF formalism, the quality of the spectral pattern depends strongly on the number of independent MD trajectories taken into account. For the model compounds tested in the gas phase, the spectral pattern was shown to adequately converge from an average of ~50 single spectra, taken from a ~200 ps simulation.

Besides the *in vacuo* calculations, Raman spectra of L-phenylalanine in solution in a QM/MM framework have been calculated using our SCC-DFTB/FTTCF implementation. Here, the same behavior concerning the intensity convergence through spectral averaging was observed as compared to the *in vacuo* calculations. However, this may not generally be true for a solute in a polar solvent. Strong intermolecular interactions, such as hydrogen bonds, could hamper an efficient sampling of phase space, making a larger number of independent trajectories necessary.

The overall spectral pattern of phenylalanine from SCC-DFTB/FTTCF calculations is in an acceptable agreement with the experiment, especially concerning the line shapes. Nevertheless, for several spectral regions, significant deviations concerning the relative band intensities were observed. Furthermore, due to well-known limitations of the SCC-DFTB approach, the C−C stretching modes in the benzene ring are strongly overestimated as indicated by a shift of ~200 $\mathrm{cm}^{-1}$ to higher wavenumbers as compared to the experiment.

Depolarization ratios estimated via the FTTCF formalism were in a good overall agreement with the other two methodologies, although SCC-DFTB/NMA compares altogether better with ratios obtained from high level BLYP aug-cc-pVTZ calculations.

Finally, in this work, we tried to estimate whether an acceptable Raman spectrum in the FTTCF framework could be obtained by replacing our SCC-DFTB/FTTCF method with a hybrid approach. Here, the calculation of molecular polarizabilities on snapshots obtained from a SCC-DFTB trajectory were performed on the B3YLP 6-31G(d) level. In fact, the spectra resulting from insufficiently sampled 40 ps trajectories unsatisfactorily reproduced NMA calculations in the frequency range below 1500 $\mathrm{cm}^{-1}$. The spectral pattern for two test cases in this region were characterized by a reduced number of intense Raman active bands compared to B3YLP 6-31G(d)/NMA. In our opinion, the hybrid approach, which incorporates high level methods for the calculation of polarizabilities, is therefore not an appropriate alternative to our SCC-DFTB/FTTCF method, since the quality of the Raman spectral pattern depends too strongly on appropriate phase space sampling. The more accurate polarizabilities obtained with B3YLP compared to SCC-DFTB cannot compensate for the effect of insufficient sampling. With the use of the highly efficient SCC-DFTB method, a sufficiently large ensemble of trajectories can be generated even for large molecules in solution, whereas for large chemical systems, the computational effort needed by a hybrid approach only allows for the generation of single short trajectories, from which the resulting Raman intensities are not representative.

**Acknowledgment.** The authors thank Rodolphe Vuilleumier for the supply of the software package for calculating effective normal modes. Financial support by the DFG Cluster of Excellence "UniCat" is acknowledged.

### References

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

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

(3) Wilson, E. B.; Decius, J. C.; Cross, P. C. *Molecular vibrations: The theory of infrared and Raman vibrational spectra*; Dover Publications: Mineola, NY, 1980; pp 11−33.

(4) McQuarrie, D. A. *Statistical Mechanics*; University Science Books: Herndon, VA, 2000; pp 467−506.

(5) Jolliffe, I. T. *Principal Component Analysis*; Springer: New York, 2002; p 299.

(6) Kitao, A.; Go, N. *Curr. Opin. Struct. Biol.* **1999**, *9*, 164.

(7) Balsera, M. A.; Wriggers, W.; Oono, Y.; Schulten, K. *J. Phys. Chem.* **1996**, *100*, 2567.

(8) Amadei, A.; Linssen, A. B.; Berendsen, H. J. *Proteins* **1993**, *17*, 412.

(9) Wheeler, R. A.; Dong, H.; Boesch, S. E. *ChemPhysChem* **2003**, *4*, 382.

(10) Gordon, R. G. *Adv. Magn. Reson.* **1968**, *3*, 1.

(11) Berens, P. H.; White, S. R.; Wilson, K. R. *J. Chem. Phys.* **1981**, *75*, 515.

(12) Gaigeot, M. P.; Vuilleumier, R.; Sprik, M.; Borgis, D. *J. Chem. Theory Comput.* **2005**, *1*, 772.

(13) Gaigeot, M. P.; Sprik, M. *J. Phys. Chem. B* **2003**, *107*, 10344.

(14) Hornicek, J.; Kaprálová, P.; Bour, P. *J. Chem. Phys.* **2007**, *127*, 4502.

(15) Pagliai, M.; Cavazzoni, C.; Cardini, G.; Erbacci, G.; Parrinello, M.; Schettino, V. *J. Chem. Phys.* **2008**, *128*, 224514.

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

(17) Lammers, S.; Meuwly, M. *J. Phys. Chem. A* **2007**, *111*, 1638.

(18) Schultheis, V.; Reichold, R.; Schropp, B.; Tavan, P. *J. Phys. Chem. B* **2008**, *112*, 12217.

(19) Asvany, O.; Kumar, P. P.; Redlich, B.; Hegemann, I.; Schlemmer, S.; Marx, D. *Science* **2005**, *309*, 1219.

(20) Kaczmarek, A.; Shiga, M.; Marx, D. *J. Phys. Chem. A* **2009**, *113*, 1985.

(21) Padma Kumar, P.; Marx, D. *Phys. Chem. Chem. Phys.* **2006**, *8*, 573.

(22) Yu, H.; Cui, Q. *J. Chem. Phys.* **2007**, *127*, 234504.

(23) Schmitz, M.; Tavan, P. *J. Chem. Phys.* **2004**, *121*, 12247.

(24) Schmitz, M.; Tavan, P. *J. Chem. Phys.* **2004**, *121*, 12233.

(25) Nonella, M.; Mathias, G.; Tavan, P. *J. Phys. Chem. A* **2003**, *107*, 8638.

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

(27) Martinez, M.; Gaigeot, M. P.; Borgis, D.; Vuilleumier, R. *J. Chem. Phys.* **2006**, *125*, 144106.

(28) Vuilleumier, R. *Mol. Phys.* **2007**, *105*, 2857.

(29) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260.

(30) Otte, N.; Scholten, M.; Thiel, W. *J. Phys. Chem. A* **2007**, *111*, 5751.

(31) Elstner, M. *Theo. Chem. Acc.* **2006**, *116*, 316.

(32) Riccardi, D.; Schaefer, P.; Yang, Y.; Yu, H.; Ghosh, N.; Prat-Resina, X.; Koenig, P.; Li, G.; Xu, D.; Guo, H.; Elstner, M.; Cui, Q. *J. Phys. Chem. B* **2006**, *110*, 6458.

(33) Witek, H. A.; Irle, S.; Morokuma, K. *J. Chem. Phys.* **2004**, *121*, 5163.

(34) Witek, H. A.; Morokuma, K. *J. Comput. Chem.* **2004**, *25*, 1858.

(35) Witek, H. A.; Morokuma, K.; Stradomska, A. *J. Chem. Phys.* **2004**, *121*, 5171.

(36) Meuwly, M.; Müller, A.; Leutwyler, S. *Phys. Chem. Chem. Phys.* **2003**, *5*, 2663.

(37) Fouqueau, A.; Meuwly, M. *J. Chem. Phys.* **2005**, *123*, 244308.

(38) Phatak, P.; Ghosh, N.; Yu, H.; Cui, Q.; Elstner, M. *Proc. Nat. Acad. Sci.* **2008**, *105*, 19672.

(39) Perdew, J. P.; Burke, K.; Wang, Y. *Phys. Rev. B* **1996**, *54*, 16533.

(40) Elstner, M. Ph.D. Thesis, University of Paderborn, Paderborn, Germany, 1998.

(41) Koch, W.; Holthausen, M. C.; Holthausen, M. C. *A chemist's guide to density functional theory*, 2nd ed.; Wiley-Vch: Weinheim, Germany, 2000; pp 177−178.

(42) Abramowitz, M.; Stegun, I. A. *Handbook of mathematical functions with formulas, graphs, mathematical tables*; Courier Dover Publications: Mineola, NY, 1964; p 884.

(43) Van Dujineveldt-Van De Rijdt, J. G. C. M.; Van Dujineveldt, F. B. *J. Mol. Struct.* **1976**, *35*, 263.

(44) McCarthy, W. J.; Lapinski, L.; Nowak, M. J.; Adamowicz, L. *J. Chem. Phys.* **1998**, *108*, 10116.

(45) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D. *J. Comput. Chem.* **1983**, *4*, 187.

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

(47) Cui, Q.; Elstner, M.; Kaxiras, E.; Frauenheim, T.; Karplus, M. *J. Phys. Chem. B* **2001**, *105*, 569.

(48) Lamoureux, G.; Roux, B. *J. Chem. Phys.* **2003**, *119*, 3025.

(49) Allen, M. P.; Tildesley, D. J. *Computer simulation of liquids*; Oxford University Press: New York, 1990; p 208.

(50) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical recipes: The art of scientific computing*, 3rd ed.; Cambridge University Press: Cambridge, U. K., 2007; p 605.

(51) Halls, M. D.; Schlegel, H. B. *J. Chem. Phys.* **1999**, *111*, 8819.

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

(53) Riley, K. E.; Op't Holt, B. T.; Merz, K. M., Jr. *J. Chem. Theory Comput.* **2007**, *3*, 407.

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

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

(56) 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 C02; Gaussian, Inc.: Wallingford, CT, 2004.

(57) Pulay, P.; Fogarazi, G.; Pang, F.; Boggs, J. E. *J. Am. Chem. Soc.* **1979**, *101*, 2550.

(58) Magdó, I.; Nemeth, K.; Mark, F.; Hildebrandt, P.; Schaffner, K. *J. Phys. Chem. A* **1999**, *103*, 289.

(59) Porezag, D.; Frauenheim, T.; Köhler, T.; Seifert, G.; Kaschner, R. *Phys. Rev. B* **1995**, *51*, 12947.

(60) Wanko, M.; Hoffmann, M.; Frauenheim, T.; Elstner, M. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 511.

(61) Małolepsza, E.; Witek, H. A.; Morokuma, K. *Chem. Phys. Lett.* **2005**, *412*, 237.

(62) Gaus, M.; Chou, C. P.; Witek, H.; Elstner, M. *J. Phys. Chem. A* **2009**, *113*, 10321.

CT900660X

JCTC Journal of Chemical Theory and Computation

# General Approach to Compute Vibrationally Resolved One-Photon Electronic Spectra

Julien Bloino,[†,‡] Malgorzata Biczysko,[‡] Fabrizio Santoro,[¶] and Vincenzo Barone*[,†]

*Scuola Normale Superiore, piazza dei Cavalieri 7, 56126 Pisa, Italy, Dipartimento di Chimica "Paolo Corradini" and CR-INSTM Village Università di Napoli Federico II, Complesso Universitario Monte S. Angelo, via Cintia, 80126 Napoli, Italy, and Istituto per i Processi Chimico Fisici, Area della Ricerca-CNR, via G. Moruzzi 1, 56124 Pisa, Italy*

Received December 16, 2009

**Abstract:** An effective time-independent approach to compute vibrationally resolved optical spectra from first principles is generalized toward the computation of one-photon electronic spectra induced by either electric or magnetic transition dipoles or by their mutual interaction. These encompass absorption, emission, and circular dichroism spectra. Additionally, the proposed computational scheme is extended to cover a broad range of approximations to evaluate vibronic transitions within both vertical and adiabatic frameworks and to be able to take into account the effects of the temperature. The presented computational tool is integrated into a general purpose computational chemistry package and offers a simple and an easy-to-use way to evaluate one-photon electronic spectra, starting from electronic structure calculations chosen according to the system under study, from fully quantum mechanical descriptions to discrete/continuum quantum mechanical/MM/polarizable continuum models.

## 1. Introduction

In a recent work,[1] we have presented a versatile procedure to compute vibrationally resolved electronic spectra, when nonadiabatic couplings are negligible, along with its integration into one of the most widely used quantum chemical packages, namely Gaussian.[2] It relies on an efficient a priori prescreening scheme[3,4] to identify the most intense transitions and to generate the spectra of medium-to-large systems[1] at a relatively nonexpensive computational cost. In a first step, the procedure was set within the adiabatic framework and was limited to one-photon absorption (OPA) and emission (OPE) transitions from the vibrational ground state of the initial electronic state, discarding the effects of the temperature. However, our general goal is to provide a robust and easy-to-use computational tool able to assist a broader range of spectroscopic studies. To this purpose, there are several issues which need to be accounted for, like electronic transitions arising from interaction between different transition dipole moments, spectral ranges encompassing more than one final electronic state, temperature effects, and anharmonicity. In the present work, we propose a generalized method able to deal with spectroscopies related either to electric or magnetic transition dipoles or to their mutual interaction. Additionally, we have introduced temperature effects and, in order to extend applicability of the approach to larger systems, also simplified models within the framework of vertical approaches. In such a way, we have completed the formulation and implementation of a general computational tool, set within the harmonic approximation and the time-independent framework, able to simulate zero- and finite-temperature electronic spectra for transitions between two electronic states, showing vibronic effects either negligible or amenable to a description within the Herzberg–Teller intensity-borrowing theory. The following discussion will be limited to OPA, OPE, and electronic circular dichroism (ECD) spectroscopies, in line with the approach[5,6] recently presented by some of us, but our method can also deal with other spectroscopic phenomena, such as the ones

* Corresponding author. E-mail: vincenzo.barone@sns.it.

[†] Scuola Normale Superiore.

[‡] Dipartimento di Chimica "Paolo Corradini" and CR-INSTM Village Università di Napoli Federico II.

[¶] Istituto per i Processi Chimico Fisici.

10.1021/ct9006772 © 2010 American Chemical Society
Published on Web 03/12/2010

issuing from magnetic dipole moments only. It is noteworthy that similar approaches have also been applied to nonresonant two-photon absorption (TPA) and circular dichroism (TPCD).[7]

Ab initio quantum approaches to the calculation of vibrationally resolved optical spectra require an a priori analysis of the relevant potential energy surfaces (PES), independently of the method adopted to compute the spectrum. However, an extensive analysis of the PES is still impractical, in most cases, and impossible for medium-to-large systems. In the latter, only a local region of the PES about a given geometry can be explored. In generic systems, an electronic excitation can involve any kind of electronic states, including neutral or ionic, and bound or dissociative ones. In addition, the initial and/or final electronic state can be subjected to strong nonadiabatic couplings, as those triggered by the existence of a conical intersection. The probability of occurrence of these couplings increases with the dimension of the system. At the current state-of-the-art, no unique rigorous method can be proposed for such a general situation. It is, therefore, convenient to clarify the reference physical model in which we developed our method: we deal with transitions between nondissociative electronic states not showing conical intersections nor strong noadiabatic effects. In such a framework, if the spectral range of interest encompasses several final electronic states, they are considered uncoupled, and the spectrum can be, therefore, calculated as a sum of the spectra of each of them. As a consequence, it is always possible to focus on a single final state at a time, and we will refer to our approach as a "single-state" approach. In this model, the most natural choice is to build both the initial and final PES, starting from a harmonic analysis at the respective equilibrium geometries. This kind of approach, often referred to as "adiabatic", treats both states on the same foot, i.e. at the same level of accuracy. It is particularly suited for high-resolution descriptions of the spectra, as it simulates correctly most of low- (high-) lying bands of the absorption (emission) spectrum, i.e., transitions to vibrational states spanning the minimum region of the final electronic state PES. Such an approach is especially important when an accurate reproduction of the fine structure of the spectrum is required, in particular, in studies related to the assignment of excited-state frequencies.[8−13] However, an important drawback of this approach is the computational cost of the geometry optimization and the frequency calculations in the excited state, which might be prohibitive for large systems. An alternative model relies on the observation that the most intense transitions are vertical, so that a correct description of the PES of the final state about the geometry of the initial state is more suited to the analysis of the region of the spectral maximum and of the broad features of the low-resolution spectrum. In fact, the latter mostly reflects the short-time dynamics of the system after an instantaneous promotion on the final state. Within the harmonic approximation, such a "vertical" approach describes the final state PES on the ground of its gradient and Hessian at the initial-state equilibrium geometry, so that it has been named VFC (vertical Franck−Condon).[14] Once the initial and final states harmonic PESs are obtained, the machinery to compute the spectra is the same for vertical and adiabatic harmonic models. Therefore, since in most cases, the computation of the excited-state Hessian is the most time-consuming step of the electronic calculation (at least when it is obtained by numerical differentiation of the gradients), the two approaches are about equivalent as far as the computational cost is concerned. From the physical point of view, as discussed above, both have their advantages and drawbacks, but in most cases, when they lead to significantly different results, the harmonic approximation itself is questionable. As an example, the VFC approach shows an increased sensitivity of the Hessian matrix to the anharmonic character of the PES. More generally, when the physical problem under investigation sensibly deviates from the reference single-state harmonic model we introduced above, no rigorous and general solution exists, and each of the adiabatic/vertical frameworks can reveal more suitable than the other for a given specific system. In the case of semirigid molecules showing conical intersections (CI), for instance, the multi-electronic-state problem can be better faced within the so-called linear vibronic coupling model (LVCM), that is based on a diabatic electronic representation and is grounded in the vertical framework. Such a model, that has been developed by the Heidelberg group in a number of seminal contributions,[15,16] is powerful and effective when harmonic approximation is suitable for describing the diabatic PES (notice that CI occurrence always makes adiabatic PES strongly anharmonic), and it has been recently adopted and generalized by Nooijen[17] to one-photon chiral spectroscopies. It is worth highlighting that when conical intersections exist in the region of the coordinate space relevant for the spectral features, attempt to use adiabatic single-state approaches may run into unsurmountable technical problems, as it is clear in the extreme case when the minimum of an adiabatic PES coincides with a CI. Albeit physically relevant, these problems may not appear evident in a vertical single-state approach, simply because the latter does not try to locate the excited-state PES minimum. It is, however, important to clarify that these situations require in principle a multi-electronic-state treatment and, even if a vertical single-state approach is technically affordable, it is not granted that it is also able to catch the main physics of the problem under investigation. From a different perspective, one may recognize that the exploration of the excited-state PES necessary to locate the excited minima can bring to light issues and problems that may simply remain unobserved in a vertical approach.

At the current state-of-the-art, the most effective implementation of LVCM for spectra calculation is based on time-dependent methods, like the multiconfigurational time-dependent Hartree (MCTDH),[18] even if time-independent (e.g., Lanczos-based) treatments are possible for limited-dimensionality problems (up to a dozen of modes) and if promising Green-function approaches have been recently proposed.[19] In the present paper, we do not deal with nonadiabatic problems, and our work is focused on the development of a robust tool for single-state harmonic spectra that can be used also by non specialists. Within this chosen framework, beyond adiabatic methods, we found it convenient to implement a simplified vertical model, by assuming that the Hessian matrix is the same in both initial and final states. Such a model, which we will refer to as vertical gradient (VG), is also known in literature as the linear coupling model (LCM)[20] (which, nonetheless, should not be confused with the multistate LVCM approach introduced above.)[15,16] In VG, only the energy gradients need to be evaluated in the final state at the equilibrium geometry of the initial one, a task much less time-consuming than the Hessian computation.

Up to now, we have discussed possible deviations from our reference physical model, arising from nonadiabatic couplings. Another issue of general relevance to be dealt with is anharmonicity. It is worthy to highlight, in fact, that its proper consideration is by far more complicated here than in vibrational spectroscopy since one has to deal contemporarily with two different electronic states. As a matter of fact, equilibrium geometries could be quite different, and this requires the description of a larger amount of the PES with the consequent problems of couplings, limits of polynomial expansions, etc. Furthermore, the normal modes of the two states can be sufficiently different to require the inclusion of large sets of coupling terms, which cause additional methodological and numerical difficulties. Moreover, minimization of coupling terms is often nonoptimal when normal modes are expressed in Cartesian coordinates, and switching to internal coordinates could be more effective.[21,22] While this is quite straightforward for small systems and/or well-defined modes, a general and automatized procedure for large systems is still lacking. This is the reason why we prefer deferring the issue of anharmonicity to further work, after we have developed a general and robust strategy at the harmonic level. Nonetheless, when anharmonicity can be considered nearly "diagonal" in terms of the normal modes of the reference state, at least two effective procedures have been proposed for semirigid systems[8] and for cases involving a single anharmonic mode,[14] respectively. The current approach incorporates a simple scheme to correct vibronic transitions for the anharmonicity in the case of semirigid systems.

In the above discussion, we have explicitly settled our reference physical model. Despite the discussed limitations, this model is able to provide a reliable interpretation of the electronic spectroscopy data for a very broad range of molecular species. On the other hand, the elaboration of a general and robust method, along with its implementation, still needs to be considered a very challenging part of the computational spectroscopy field. To this purpose, two main and strictly correlated issues must be taken into proper account, the former concerns the reliability of the electronic description, while the latter is more focused on the vibrational problem. More specifically, the first issue related to the evaluation of properties of molecular systems in their excited electronic states affects the reliability of computational spectroscopy studies. In fact, until recently, computations of vibronic spectra have been limited to small molecules, but recent developments in electronic structure theory for excited states within the time-dependent density functional theory (TD-DFT)[23,24] and the resolution of the identity approximation of coupled cluster theory (RI-CC2)[25] have paved the route toward the simulation of spectra for significantly larger systems. In this respect, a pivotal role can be played by hybrid DFT/CC approaches where the most expensive computationally geometry optimization and Hessian computations are performed by DFT with medium-size basis sets, while energies and/or other properties can be computed with more accurate quantum mechanics (QM) approaches.[8] However, when larger systems are to be studied, besides the QM treatment, a second computational challenge arises from the inclusion of vibrational contributions. Indeed, the number of vibrational states to be taken into account increases steeply with the dimension of the system and with the spectral width, but most of the possible vibronic transitions do not contribute significantly to the spectrum and can be safely neglected. Therefore, wise and effective selection criteria to individuate a priori the most relevant vibronic transitions within the dense bath of possible final states can make feasible the calculations of spectrum lineshapes for macromolecular systems. Several schemes have been proposed[3,26-30] ranging from the simplest approach, based solely on the energy window of the spectrum,[26,27] up to rigorous prescreening techniques, based on analytically derived sum rules.[30] In order to maximize the efficiency of calculations and to deal with large systems, it is necessary to adopt a fast, a priori selection scheme of general applicability for a variety of different systems that is able to correctly choose all the non-negligible transitions. We have recently derived a general and robust tool rooted into a method recently introduced by some of us[3,29] in the frame of harmonic approximation, which has been proven to provide very accurate spectra of medium-to-large systems with a limited computational cost.[8] Moreover, in the implementation,[1] particular care has been taken to avoid the introduction of any built-in restriction, be it for the number of allowed quanta in a single mode, for the number of simultaneously excited vibrations, or for the spectrum energy range. In this paper, we extend the applicability of our tool to other spectroscopic phenomena, temperature effects, and even larger systems (through the VG model). Thanks to the new available features and to the full integration into the Gaussian package, the one-photon electronic spectra for a wide variety of cases can be treated in a fully automatic way easily accessible also to nonspecialists. In this work, the developed tool is applied to a variety of different systems and problems, such as simulations of OPA, OPE, and ECD spectra with and without temperature effects and both in the gas phase and in solution (where the solvent is described within polarizable continuum models).

The paper is organized as follows: Section 2 describes the theoretical frame of the generalized approach to compute one-photon electronic spectra, along with the details of the current implementation. Computational models, which have been applied to the determination of structures, forces, vibrations, and energies to provide the information necessary for spectra calculation, are gathered in Section 3. The developed method gave us the opportunity to extensively validate the DFT/N07 model for the computations of ECD spectra, and results are reported in Section 4. Finally, the simulated vibrationally resolved one-photon electronic spectra are gathered in Section 5. Aspects of simulated spectra convergence are discussed in Sections 5.1 and 5.2 on the example of the well studied $S_2 \leftarrow S_0$ one-photon ECD spectrum of (*R*)-(+)-3 methylcyclopentanone (R3MCP), which is followed by two illustrative examples of direct application of our integrated approach to larger systems: the UV−vis absorption spectrum in the 250−700 nm range of chlorophyll a1 (Section 5.3) in methanol solution, and the ECD spectrum of $\pi^* \leftarrow n$ transition of (*Z*)-8-methoxy-4-cyclooctenone (MCO) (Section 5.4).

## 2. Computation of One-Photon Electronic Spectra

**2.1. Generalized Model.** In order to deal with different one-photon electronic spectroscopies, the procedure described in our previous work[1] has been redesigned to be as general as possible. Additionally, it has been modified to provide quantities easier to compare with their experimental counterparts. In this context, the intensity of OPA, fluorescence, OPE and ECD can be expressed by the same equation:

$$I = \alpha\omega^{\beta} \sum_{m} \sum_{n} \rho_{\gamma}[\boldsymbol{d}_{\mathrm{A}mn} \cdot \boldsymbol{d}^{*}_{\mathrm{B}mn}]\delta(\omega_n - \omega_m - \omega) \quad (1)$$

where the symbol "*" is used to represent the conjugate of $\boldsymbol{d}_{\mathrm{B}mn}$, and $\delta$ is the Dirac function.

The quantity $I$ is evaluated by taking into account all transition probabilities, represented by the double summation over all possible initial vibronic states $m$, weighted by the probability of the molecule to be in this initial state given by the Boltzmann population $\rho_\gamma$ and by the final vibronic states $n$.

The intensity for OPA, OPE, or ECD is given by replacing $I$, $\alpha$, $\beta$, $\gamma$, $\boldsymbol{d}_{\mathrm{A}mn}$, and $\boldsymbol{d}_{\mathrm{B}mn}$ with the values given in the list below:

$$\text{OPA: } I = \varepsilon(\omega),\ \alpha = \frac{10\pi\mathcal{N}_{\mathrm{A}}}{3\varepsilon_0 \ln(10)\hbar c},\ \beta = 1,\ \gamma = m,$$
$$\boldsymbol{d}_{\mathrm{A}mn} = \boldsymbol{d}_{\mathrm{B}mn} = \boldsymbol{\mu}_{mn}$$

$$\text{OPE: } I = I_{em}/N_n,\ \alpha = \frac{2\mathcal{N}_{\mathrm{A}}}{3\varepsilon_0 c^3},\ \beta = 4,\ \gamma = n,$$
$$\boldsymbol{d}_{\mathrm{A}mn} = \boldsymbol{d}_{\mathrm{B}mn} = \boldsymbol{\mu}_{mn}$$

$$\text{ECD: } I = \Delta\varepsilon(\omega),\ \alpha = \frac{40\mathcal{N}_{\mathrm{A}}\pi}{3\varepsilon_0 \ln(10)\hbar c^2},\ \beta = 1,\ \gamma = m,$$
$$\boldsymbol{d}_{\mathrm{A}mn} = \boldsymbol{\mu}_{mn},\ \boldsymbol{d}_{\mathrm{B}mn} = \mathcal{I}(\boldsymbol{m}_{mn})$$

where $\varepsilon(\omega)$ is the molar absorption coefficient for a given angular frequency $\omega$, $\Delta\varepsilon(\omega)$ is the difference (referred to as anisotropy) between the molar absorption coefficients $\varepsilon^-$ and $\varepsilon^+$, relative to the left and right circularly polarized light, respectively. For OPE, $I_{em}/N_n$ is the energy emitted by one mole per second. It is noteworthy that phosphorescence spectra, i.e., emission from electronic states characterized by different spin multiplicity with respect to the ground electronic state, depend on spin−orbit couplings that are not always available. In this case, only Franck−Condon calculations are performed (see below for details), assuming the transition is fully allowed with the length of the electric transition dipole moment set to 1 au.

Finally, $\mathcal{N}_{\mathrm{A}}$ is the Avogadro constant, $\varepsilon_0$ is the vacuum permittivity, $\boldsymbol{\mu}_{mn}$ is the electric transition dipole moment between the vibronic states $m$ and $n$, and $\mathcal{I}(\boldsymbol{m}_{mn})$ is the imaginary part of the magnetic transition dipole moment between the vibronic states $m$ and $n$, $\boldsymbol{m}_{mn}$. A more detailed description of the calculation of $I$ for OPA, OPE, and ECD is available in ref 31.

We will use from now on the symbol $\boldsymbol{d}_{\mathrm{X}}$ to represent indifferently $\boldsymbol{d}_{\mathrm{A}}$ and $\boldsymbol{d}_{\mathrm{B}}$. We define the integral $\boldsymbol{d}_{\mathrm{X}mn}$ and its conjugate as

$$\boldsymbol{d}_{\mathrm{X}mn} = \langle \Psi_m | \boldsymbol{d}_{\mathrm{X}} | \Psi_n \rangle \quad (2)$$

$$\boldsymbol{d}^{*}_{\mathrm{X}mn} = \langle \Psi_n | \boldsymbol{d}^{*}_{\mathrm{X}} | \Psi_m \rangle \quad (3)$$

where $\Psi_m$ and $\Psi_n$ are the molecular wave functions of the vibronic states $m$ and $n$, respectively.

From eq 1, one can see that the knowledge of $I$ is bound to the evaluation of $\boldsymbol{d}_{\mathrm{A}mn} \cdot \boldsymbol{d}_{\mathrm{B}mn}{}^{*}$.

In the framework of the Born−Oppenheimer approximation, the molecular wave function $\Psi$ can be written as a product of its electronic and nuclear components, $\psi_{\mathrm{e}}$ and $\psi_{\mathrm{N}}$, respectively. For readability, the spectroscopic convention will be used to designate the wave functions, a single quote (′) representing the lower vibronic state in energy $m$ and a double quote (″) the higher state, $n$.

$$\langle \Psi' | \boldsymbol{d}_{\mathrm{X}} | \Psi'' \rangle = \langle \psi'_{\mathrm{e}} \psi'_{\mathrm{N}} | \boldsymbol{d}^{\mathrm{e}}_{\mathrm{X}} | \psi''_{\mathrm{e}} \psi''_{\mathrm{N}} \rangle + \langle \psi'_{\mathrm{e}} \psi'_{\mathrm{N}} | \boldsymbol{d}^{\mathrm{N}}_{\mathrm{X}} | \psi''_{\mathrm{e}} \psi''_{\mathrm{N}} \rangle \quad (4)$$

Because of the orthogonality of the electronic wave functions, the second term in the right-hand side (rhs) of the previous equation is null, so that:

$$\langle \Psi' | \boldsymbol{d}_{\mathrm{X}} | \Psi'' \rangle = \langle \psi'_{\mathrm{e}} \psi'_{\mathrm{N}} | \boldsymbol{d}^{\mathrm{e}}_{\mathrm{X}} | \psi''_{\mathrm{e}} \psi''_{\mathrm{N}} \rangle = \langle \psi'_{\mathrm{N}} | \boldsymbol{d}^{\mathrm{e}}_{\mathrm{X}mn} | \psi''_{\mathrm{N}} \rangle \quad (5)$$

Finally, assuming that the Eckart conditions[32] are met, it is possible to separate, with a good approximation, the nuclear wave function into translational, rotational, and vibrational terms. As we are interested in the vibrational contribution in radiative transitions, we discard the translational and rotational wave functions:

$$\langle \Psi' | \boldsymbol{d}_{\mathrm{X}} | \Psi'' \rangle = \langle \psi'_{v} | \boldsymbol{d}^{\mathrm{e}}_{\mathrm{X}mn} | \psi''_{v} \rangle$$

More precisely, the intensity in eq 1 is obtained after an orientational averaging that assumes freely rotating molecules. The electric and magnetic transition dipole moments, $\boldsymbol{\mu}^{e}_{mn}$ and $\boldsymbol{m}^{e}_{mn}$ respectively, are, in general, unknown functions of the vibrational coordinates so that the transition dipole moment $\boldsymbol{d}^{\mathrm{e}}_{\mathrm{X}mn}$, which could represent either $\boldsymbol{\mu}^{\mathrm{e}}_{mn}$ or $\mathcal{I}(\boldsymbol{m}^{\mathrm{e}}_{mn})$, must be approximated. The most common approximation, stated by Franck[33] and formalized by Condon,[34] assumes that the transition takes place in such a short time that the position of the nuclei remains almost unchanged and that the transition dipole can be considered as constant. While this approximation is fairly good when the transition is fully allowed and the minima of the potential energy surfaces of the initial and final electronic states are almost vertically to each other, it shows serious limitations when the transition is weakly allowed or dipole forbidden. The limitation is even more strongly felt for ECD where the product $\boldsymbol{\mu}_{mn}\cdot\mathscr{I}(\boldsymbol{m}_{mn})$ can be almost negligible, even if the transition is strongly allowed whenever the electric and magnetic moments are nearly orthogonal. An early extension to the Franck−Condon principle was proposed by Herzberg and Teller[35] and accounted for a linear variation of the transition dipole moment, with respect to the normal coordinates of the initial state, $\boldsymbol{Q}'$ or $\boldsymbol{Q}''$, depending if absorption or emission is considered. It must be recognized, however, that, once the most suitable physical model has been chosen and the proper quantities obtained from electronic calculations (e.g., from the expansion around the initial-state geometry), from the mathematical point of view, the transition dipole moment $\boldsymbol{d}^{\mathrm{e}}_{\mathrm{X}mn}$ can be equally well expressed in a Taylor expansion around the equilibrium geometry of each electronic state. For convenience, with respect to the formulation of the overlap integrals in the rest of the document, the higher state in energy is chosen as the reference. The Taylor expansion about the equilibrium geometry of this state, $\boldsymbol{Q}''_0 = \boldsymbol{0}$, is

$$\boldsymbol{d}^{\mathrm{e}}_{\mathrm{X}mn}(\boldsymbol{Q}'') \approx \boldsymbol{d}^{\mathrm{e}}_{\mathrm{X}mn}(\boldsymbol{Q}''_0) + \sum_{i=1}^{N}\left(\frac{\partial \boldsymbol{d}^{\mathrm{e}}_{\mathrm{X}mn}}{\partial \mathrm{Q}''_i}\right)_0 \mathrm{Q}''_i + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{\partial^2 \boldsymbol{d}^{\mathrm{e}}_{\mathrm{X}mn}}{\partial \mathrm{Q}''_i\partial \mathrm{Q}''_j}\right)_0 \mathrm{Q}''_i\mathrm{Q}''_j + ... \qquad (6)$$

where $N$ is the number of normal modes.

The FC approximation corresponds to the first term in the rhs of eq 6, HT to the second one and Franck−Condon Herzberg−Teller (FCHT) to both terms. The remaining terms in eq 6 will not be taken into account in the following discussion. From a vibronic point of view, the HT term introduces an intensity borrowing effect due to the interaction of the states involved in the electronic transition with other closely lying electronic states, with which they mix upon small displacements along the normal coordinates. When an electronic multistate diabatic representation is adopted, like in the linear coupling vibronic model cited in Section 1 (the Introduction), Herzberg−Teller effects are usually less important since diabatic states are defined in order to be ideally independent of the coordinates. In those cases, intensity borrowing mechanisms are explicitly introduced by the coupling among the different diabatic states. However, when the interacting states are sufficiently separated in energy and when the interaction is weak, traditional HT treatment allows to account for the main borrowing effects, keeping the simplicity of an adiabatic single-state description.

At this stage, it is useful to digress from the theoretical background and to take a deeper look into the computational implementation. While our method is fully general and can be applied starting from data obtained with any electronic model, it is oriented toward large molecules. It must be recognized that the state of the art methods rooted to TD-DFT provide probably the most effective route to treat such systems. In this approach, the excited-state Hessian, required for harmonic analysis, is computed by numerical differentiation of the energy gradient at the equilibrium geometry of the excited state. During this calculation, it is also possible to obtain, at no additional cost, the numerical derivatives of the electric and magnetic transition dipole moments. Therefore, by default, our implementation uses these sets of data for HT and FCHT calculations. It must be remembered, however, that, as previously discussed in ref 4, Herzberg and Teller considered in their original work[35] the initial state as the state of reference for the linear variation of the electronic transition dipole moment (we call it $^{i}$HT), in line with the FC principle. Therefore, while fully agreeing with their original approach for emission, our default computational implementation slightly differs from their original proposal for absorption processes, adopting the final-state equilibrium geometry reference ($^{f}$HT). It should be noted that nondefault choices of proper keywords allow a calculation along the original $^{i}$HT formulation, to be performed also for absorption processes. However, we point out that when the $^{i}$HT and $^{f}$HT approaches lead to substantial different results, the common linear approximation must probably be questioned.

Regarding the vertical gradient (VG) approach, it is possible to combine it with the FC and the FCHT approximations, as often done in the literature.[20] Nonetheless, at least at the TD-DFT level, the effort required for the numerical differentiation of the transition dipole derivatives neutralizes the computational convenience of the VG model, since at the same cost the excited state normal modes also can be obtained. Hence, in the following, the VG is only adopted in combination with the FC approximation.

Using eq 6, the transition dipole moment integral $\boldsymbol{d}_{\mathrm{X}mn}$ is given by the relation

$$\langle\Psi_m|\boldsymbol{d}_x|\Psi_n\rangle \approx \boldsymbol{d}^{\mathrm{e}}_{\mathrm{X}mn}(\boldsymbol{Q}''_0)\langle\psi'_v|\psi''_v\rangle + \sum_{i=1}^{N}\left(\frac{\partial \boldsymbol{d}^{\mathrm{e}}_{\mathrm{X}mn}}{\partial \mathrm{Q}''_i}\right)_0\langle\psi'_v|\mathrm{Q}''_i|\psi''_v\rangle \qquad (7)$$

The overlap integral $\langle\psi'_v|\psi''_v\rangle$ is also referred as the Franck−Condon integral.

While the treatment can be done at the anharmonic level,[36,37] our current implementation treats the calculation of the Franck−Condon integrals at the harmonic level. With this approximation, the multidimensional wave function $\psi_v$ can be written as a product of one-dimensional vibrational wave functions $\psi_{v_i}$,

$$|\psi_v\rangle = \prod_{i=1}^{N}|\psi_{v_i}\rangle \qquad (8)$$

For purposes of compactness, the convenient Dirac notation will be adopted from now on, $|\psi_v\rangle = |\boldsymbol{v}\rangle$ where $v$ represents the vector of quantum numbers $v_i$ for each vibrational mode $i$.

Using second quantization, the second term in the rhs of eq 7, which depends on the normal coordinates $\boldsymbol{Q}''$, can be reformulated as

$$\langle\boldsymbol{v}'|\mathrm{Q}''_i|\boldsymbol{v}''\rangle = \sqrt{\frac{\hbar}{2\omega''_i}}[\sqrt{v''_i}\langle\boldsymbol{v}'|\boldsymbol{v}'' - \boldsymbol{1}''_i\rangle + \sqrt{v''_i + 1}\langle\boldsymbol{v}'|\boldsymbol{v}'' + \boldsymbol{1}''_i\rangle] \qquad (9)$$

***Table 1.*** Ab Initio Computations Required to Generate the Input Data for the Simulation Vibrationally Resolved Electronic Spectra with the VG, AS, AFC, and AFCHT Models

| computation | VG[a] | AS[a] | AFC | AFCHT |
|---|---|---|---|---|
| initial state | | | | |
| Cartesian coordinates of the atoms (equilibrium structure) | x | x | x | x |
| energy at the minimum of the PES (equilibrium geometry) | x | x | x | x |
| frequencies | x | x | x | x |
| normal modes, expressed by the atomic displacements | x | x | x | x |
| final state | | | | |
| Cartesian coordinates of the atoms at the minimum of the PES (equilibrium structure) | | x | x | x |
| energy at the equilibrium geometry of the initial state | x | | | |
| energy at the minimum of the PES (equilibrium geometry) | | x | x | x |
| forces at the equilibrium geometry of the initial state | x | | | |
| frequencies | | | x | x |
| normal modes, expressed by the atomic displacements | | | x | x |
| general | | | | |
| atomic masses | x | x | x | x |
| transition dipole moments | x | x | x | x |
| derivatives of the transition dipole moments | | | | x |

[a] For the VG and AS approaches, it is assumed that the FC approximation is used. See text for the details.

Introducing eq 9 in eq 7, we obtain the following Taylor expansion for the transition dipole moment:

$$\langle\Psi_m|\boldsymbol{d}_x|\Psi_n\rangle \approx \boldsymbol{d}^{\mathrm{e}}_{\mathrm{X}mn}(\boldsymbol{Q}''_0)\langle\boldsymbol{v}'|\boldsymbol{v}''\rangle + \sum_{i=1}^{N}\left(\frac{\partial\boldsymbol{d}^{\mathrm{e}}_{\mathrm{X}mn}}{\partial \mathrm{Q}''_i}\right)_0\sqrt{\frac{\hbar}{2\omega''_i}}[\sqrt{v''_i}\langle\boldsymbol{v}'|\boldsymbol{v}''-1''_i\rangle + \sqrt{v''_i+1}\langle\boldsymbol{v}'|\boldsymbol{v}''+1''_i\rangle] \tag{10}$$

A major issue to calculate the integrals in eq 10 arises from the fact that each vibrational wave function is expressed in a different set of normal coordinates. This problem can be overcome by the linear transformation proposed by Duschinsky[38] to express the normal coordinates of one state with respect to the other's:

$$\boldsymbol{Q}^i = \boldsymbol{J}\boldsymbol{Q}^f + \boldsymbol{K} \tag{11}$$

where $\boldsymbol{J}$ is the so-called Duschinsky matrix and represents the mixing of the normal modes during the transition, and $\boldsymbol{K}$ is the shift vector of the normal modes between the initial and final states. $\boldsymbol{Q}^i$ represents the normal coordinates of the initial state, and $\boldsymbol{Q}^f$ represents those of the final one. In case of OPA or ECD spectroscopies, $\boldsymbol{Q}^i = \boldsymbol{Q}'$ and $\boldsymbol{Q}^f = \boldsymbol{Q}''$, and in OPE, $\boldsymbol{Q}^i = \boldsymbol{Q}''$ and $\boldsymbol{Q}^f = \boldsymbol{Q}'$.

The rotation or Duschinsky matrix is given by

$$\boldsymbol{J} = (\boldsymbol{L}^i)^{-1}\boldsymbol{L}^f \tag{12}$$

where $\boldsymbol{L}^i$ and $\boldsymbol{L}^f$ are the transformation matrices from mass-weighted Cartesian coordinates to normal coordinates of the initial and final states, respectively.

When considering an adiabatic model, the shift vector is given by the difference of geometries between the final and initial states:

$$\boldsymbol{K} = (\boldsymbol{L}^i)^{-1}\boldsymbol{M}^{1/2}\Delta\boldsymbol{X} \tag{13}$$

where $\boldsymbol{M}$ is the diagonal matrix of atomic masses, and $\Delta\boldsymbol{X} = \boldsymbol{X}_0{}^f - \boldsymbol{X}_0{}^i$ is a vector representing the shift of nuclear Cartesian coordinates between the initial ($\boldsymbol{X}^i$) and final ($\boldsymbol{X}^f$) states.

For "vertical" models, the structure remains unchanged, but information about the shift of the normal modes between the initial and final states are obtained from the energy gradient in the final state. Since our treatment of the vertical models will be limited here to VG, we report only the shift used for this approximation:

$$\boldsymbol{K} = -\{\boldsymbol{\Omega}^i\}^{-2}[(\boldsymbol{L}^i)^{-1}\boldsymbol{M}^{-1/2}\boldsymbol{g}^f] \tag{14}$$

where $\boldsymbol{\Omega}^i$ is the diagonal matrix of the harmonic frequencies $\boldsymbol{\omega}$ of the initial state, and $\boldsymbol{g}^f$ is the final state's energy gradient in Cartesian coordinates.

It should be noted that in this model, the Hessian matrix of the final state is assumed to be the same as in the initial state. As a consequence, we have $\boldsymbol{L}^i = \boldsymbol{L}^f$ so that $\boldsymbol{J}$ is the identity matrix $\boldsymbol{I}$. Moreover, it is assumed that the final-state frequencies are identical to the initial-state ones. Based on these approximations, we can also define an "adiabatic" model, that we will refer to as adiabatic shift, where $\boldsymbol{J}$ is replaced by the identity matrix while the shift vector $\boldsymbol{K}$ is kept unchanged with respect to the value given in eq 13. Notice that, at variance with the adiabatic models where the shift vector only depends on structural parameters (the two equilibrium geometries), thus representing a true displacement, in VG $\boldsymbol{K}$ not only depends on the vibrational frequencies of the initial state but it is also sensitive to anharmonicities (through the gradient), and it must be considered, therefore, a sort of "effective displacement".

The overlap integrals can then be evaluated analytically[39−44] or recursively.[45−48] While the former method allows straightforward calculations and avoids possible error propagation, it suffers from a quickly growing complexity and a lack of versatility when dealing with medium-to-large systems. As a consequence, the more general-purpose recursive approach presented by Ruhoff[47] and based on the generating functions of Sharp and Rosenstock[39] has been implemented:

$$\langle \boldsymbol{v}'|\boldsymbol{v}''\rangle = \frac{1}{\sqrt{2v_i'}}\left[\mathrm{B}_i\langle \boldsymbol{v}' - 1_i'|\boldsymbol{v}''\rangle + \sqrt{2(v_i' - 1)}\mathrm{A}_{ii}\langle \boldsymbol{v}' - 2_i'|\boldsymbol{v}''\rangle + \sum_{j=1,j\neq i}^{N}\sqrt{2v_j'}\mathrm{A}_{ij}\langle \boldsymbol{v}' - 1_i' - 1_j'|\boldsymbol{v}''\rangle + \sum_{j=1}^{N}\sqrt{\frac{v_j''}{2}}\mathrm{E}_{ji}\langle \boldsymbol{v}' - 1_i'|\boldsymbol{v}'' - 1_j''\rangle\right] \tag{15}$$

and

$$\langle \boldsymbol{v}'|\boldsymbol{v}''\rangle = \frac{1}{\sqrt{2v_i''}}\left[\mathrm{D}_i\langle \boldsymbol{v}'|\boldsymbol{v}'' - 1_i''\rangle + \sqrt{2(v_i'' - 1)}\mathrm{C}_{ii}\langle \boldsymbol{v}'|\boldsymbol{v}'' - 2_i''\rangle + \sum_{j=1,j\neq i}^{N}\sqrt{2v_j''}\mathrm{C}_{ij}\langle \boldsymbol{v}'|\boldsymbol{v}'' - 1_i'' - 1_j''\rangle + \sum_{j=1}^{N}\sqrt{\frac{v_j'}{2}}\mathrm{E}_{ij}\langle \boldsymbol{v}' - 1_j'|\boldsymbol{v}'' - 1_i''\rangle\right] \tag{16}$$

where the matrices and vectors **A**, **B**, **C**, **D**, and **E** are the Sharp and Rosenstock matrices.

**2.2. An Efficient A Priori Method to Generate One-Photon Electronic Spectra.** Once the final electronic state is individuated, no selection rule (beyond those based on the symmetry of the vibrational wave functions) exists to effectively limit the huge number of possibly bound final states to be taken into account in sizable molecules. However, transitions to most of them have a negligible intensity. Based on this observation, a prescreening method[3,4] is used to select a priori the most intense transitions. It relies on a categorization of the latter with respect to the number of simultaneously excited modes in the final state, called classes. For instance, class 1 ($\mathcal{C}_1$) represents all transitions to final vibrational states with a single excited mode $i$, $\langle \boldsymbol{v}'|\mathbf{0}'' + v_i''\rangle$, and class 0 contains the overlap integral between the vibrational ground states, $\langle \mathbf{0}'|\mathbf{0}''\rangle$. Based on this division, the overlap integrals in classes 1 and 2 are used as reference data to prescreen those to compute in each "higher" class, each class being calculated one after the other, increasing the number of excited modes in the final state. The advantages of using the overlap integrals of these classes are two-fold. The first one is that these integrals are computationally cheap and are generated quickly even in the case of large molecules. The second interest lies in the information provided by the reference data. We present here a generalized approach able to handle OPA, OPE, and ECD spectroscopies, but which could be extended easily to additional kinds of spectroscopy.

Depending if only the zeroth order of the Taylor expansion given in eq 6 or higher orders are considered, two or three data sets are required. This number is doubled when temperature is taken into account. The first general set, $F_{\mathcal{C}_1}$, is defined during the calculation of the transtions from class 1. Its elements are defined as

$$F_{\mathcal{C}_1}(i, v_i'') = \left|\frac{1}{\sqrt{2v_i''}}\left[\mathrm{D}_i\langle \mathbf{0}'|\mathbf{0}'' + v_i'' - 1_i''\rangle + \sqrt{2(v_i'' - 1)}\mathrm{C}_{ii}\langle \mathbf{0}'|\mathbf{0}'' + v_i'' - 2_i''\rangle\right]\right|^2 \tag{17}$$

where the factors $\mathrm{C}_{ii}$ and $\mathrm{D}_i$ give, respectively, information on the effect of the shifts in equilibrium positions and the frequency changes on the overlap integrals of overtones and, more precisely, on the vibrational progression of mode $i$.

The second set, $F_{\mathcal{C}_2}$, is obtained in class 2 and gives information about the Duschinsky mixing of the normal modes. It contains all combinations of modes $i$ and $j$ but only considering the cases of an equal number of quanta for both modes ($v_i'' = v_j''$):

$$F_{\mathcal{C}_2}(i,j, v_i'' = v_j'') = |\langle \mathbf{0}'|\mathbf{0}'' + v_i'' + v_j''\rangle|^2 - \frac{F_{\mathcal{C}_1}(i, v_i'') \times F_{\mathcal{C}_1}(j, v_i'')}{|\langle \mathbf{0}'|\mathbf{0}''\rangle|^2} \tag{18}$$

For Herzberg−Teller calculations (HT or FCHT approximations), a second data set is extracted from class 1, $H_{\mathcal{C}_1}$, which stores an upper-bound estimation of the square of the pure Herzberg−Teller contribution for a given mode $i$ and the corresponding transition $\langle \mathbf{0}'|\mathbf{0}'' + v_i''\rangle$:

$$\begin{aligned} H_{\mathcal{C}_1}(i, v_i'') \quad = \quad & \sum_{\tau=x,y,z}\left|\left(\frac{\partial \boldsymbol{d}^{\mathrm{e}}_{\mathrm{A}mn}(\tau)}{\partial \mathrm{Q}_i''}\right)_0\right|\left|\left(\frac{\partial \boldsymbol{d}^{\mathrm{e}*}_{\mathrm{B}mn}(\tau)}{\partial \mathrm{Q}_i''}\right)_0\right| \\ & \times \quad \left|\sqrt{\frac{\hbar}{2\omega_i''}}[\sqrt{v_i''}\langle \mathbf{0}'|\mathbf{0}'' + v_i'' - 1_i''\rangle + \sqrt{v_i'' + 1}\langle \mathbf{0}'|\mathbf{0} + v_i'' + 1_i''\rangle]\right|^2 \end{aligned} \tag{19}$$

with the summation over each Cartesian coordinate of the transition dipole moments $\boldsymbol{d}^{\mathrm{e}}_{\mathrm{A}mn}$ and $\boldsymbol{d}^{\mathrm{e}}_{\mathrm{B}mn}$.

The method to choose the maximum quantum number of each mode has been extensively presented in previous references.[3,4]

When considering temperature, an additional difficulty lies in the choice of the starting vibrational states in the initial electronic state. An evident way to limit the treatment is to use a threshold on the Boltzmann population of each vibrational state. In practice, this threshold is set with respect to the population of the ground state. Similarly to the final states, a division in classes is performed for the initial states. For each class, a set is defined by the initial states sharing the same simultaneously excited modes, so that they differ only by the quantum numbers of these modes, and each set (in previous papers named "mother states")[4,29] is treated separately. The a priori selection of the most intense transitions requires two additional (or three in the case of the Herzberg−Teller approximation) data sets, $F^{\mathrm{T}}_{\mathcal{C}_1}$ and $F^{\mathrm{T}}_{\mathcal{C}_2}$ (and $H^{\mathrm{T}}_{\mathcal{C}_1}$), which are equivalent to $F_{\mathcal{C}_1}$ and $F_{\mathcal{C}_2}$ (and $H_{\mathcal{C}_1}$) for the case of finite temperature. These sets are defined for the highest-energy initial state for each set:

$$F^{\mathrm{T}}_{\mathcal{C}_1}(i, v''_i) = \left|\frac{1}{\sqrt{2v''_i}}\left[\mathrm{D}_i\langle \boldsymbol{v}'|\mathbf{0}'' + v''_i - 1''_i\rangle + \sqrt{2(v''_i - 1)}\mathrm{C}_{ii}\langle \boldsymbol{v}'|\mathbf{0}'' + v''_i - 2''_i\rangle + \sum_{j=1}^{N}\sqrt{\frac{v'_j}{2}}\mathrm{E}_{ij}\langle \boldsymbol{v}' - 1'_j|\mathbf{0}'' + v''_i - 1''_i\rangle\right]\right|^2 \quad (20)$$

$$F^{\mathrm{T}}_{\mathcal{C}_2}(i, j, v''_i = v''_j) = |\langle \boldsymbol{v}'|\mathbf{0}'' + v''_i + v''_i\rangle|^2 - \frac{F^{\mathrm{T}}_{\mathcal{C}_1}(i, v''_i) \times F^{\mathrm{T}}_{\mathcal{C}_1}(j, v''_i)}{|\langle \boldsymbol{v}'|\mathbf{0}''\rangle|^2} \quad (21)$$

$$H^{\mathrm{T}}_{\mathcal{C}_1}(i, v''_i) \;=\; \sum_{\tau=x,y,z}\left|\left(\frac{\partial \boldsymbol{d}^{\mathrm{e}}_{\mathrm{A}mn}(\tau)}{\partial \mathrm{Q}''_i}\right)_0\right|\left|\left(\frac{\partial \boldsymbol{d}^{\mathrm{e}}_{\mathrm{B}mn}{}^*(\tau)}{\partial \mathrm{Q}''_i}\right)_0\right| \times \left|\sqrt{\frac{\hbar}{2\omega''_i}}\left[\sqrt{v''_i}\langle \boldsymbol{v}'|\mathbf{0}'' + v''_i - 1''_i\rangle + \sqrt{v''_i + 1}\langle \boldsymbol{v}'|\mathbf{0} + v''_i + 1''_i\rangle\right]\right|^2 \quad (22)$$

Details of the prescreening method when taking into account the temperature can be found in ref 29.

When choosing a priori the transitions to compute, it is necessary to control the reliability of this prescreening. This is done by comparing the total intensity $I^{\mathrm{calc}}$ obtained by the addition of all the transitions taken into account to the expected intensity $I^{\mathrm{tot}}$ calculated using analytic sum rules.

The method to compute the "analytic" total intensities has been presented in previous papers, approximating the electronic transition dipole moment in a Taylor series up to the second order.[1,4] For completeness, we report the generalized calculation of the spectrum convergence ($I^{\mathrm{calc}}/I^{\mathrm{tot}}$) for any couple of dipoles $\boldsymbol{d}_{\mathrm{A}}$ and $\boldsymbol{d}_{\mathrm{B}}$ in the Supporting Information of the present contribution. Starting from eq 1, the total intensity can be defined as

$$I^{\mathrm{tot}} = \sum_i \rho^i \langle \boldsymbol{v}^i|\boldsymbol{d}^{\mathrm{e}}_{\mathrm{A}mn} \cdot \boldsymbol{d}^{\mathrm{e}}_{\mathrm{B}mn}{}^*|\boldsymbol{v}^i\rangle \quad (23)$$

where the summation is performed over all the initial vibrational states $i$ and $\rho^i$ is the Boltzmann population of the initial state $|\boldsymbol{v}^i\rangle$. For OPA and ECD, the initial state is the lowest in energy, while for OPE, it is the highest one.

With these definitions, Supporting Information shows that the spectrum convergence, $\varsigma$, is given by the relation:

$$\varsigma = \frac{\sum_i \sum_f \rho_{v^i}\langle \boldsymbol{v}^i|\boldsymbol{d}^{\mathrm{e}}_{\mathrm{A}mn}|v^f\rangle\langle v^f|\boldsymbol{d}^{\mathrm{e}}_{\mathrm{B}mn}{}^*|\boldsymbol{v}^i\rangle}{\sum_i \rho^i\left[\boldsymbol{d}^{\mathrm{e}}_{\mathrm{A}mn}(\boldsymbol{Q}^i_0) \cdot \boldsymbol{d}^{\mathrm{e}}_{\mathrm{B}mn}{}^*(\boldsymbol{Q}^i_0) + \sum_{k=1}^{N}\left(\frac{\partial \boldsymbol{d}^{\mathrm{e}}_{\mathrm{A}mn}}{\partial \mathrm{Q}^i_k}\right)_0\left(\frac{\partial \boldsymbol{d}^{\mathrm{e}}_{\mathrm{B}mn}{}^*}{\partial \mathrm{Q}^i_k}\right)_0\frac{\hbar}{2\omega'_k}(2v'_k + 1)\right]} \quad (24)$$

where $|\boldsymbol{v}^f\rangle$ represents a vibrational final state.

**2.3. Implementation.** A wide range of approaches to compute vibrationally resolved electronic spectra, for cases where the involved electronic states can be considered not-coupled, has been presented, which differ by their conceptual approach to the transition, vertical, or adiabatic as well as the level of approximation of the respective PESs of the initial and final states, such for example adiabatic FC and adiabatic shift (AS). Additionally, various approximations of the transition dipole moments, FC, FCHT, or HT can be applied. The acronyms AFC, AFCHT, and AHT will be used to refer to the adiabatic model with the harmonic representation of the PES of each electronic state computed at its equilibrium geometry, differing only by the approximation of the transition dipole moment in eq 6, respectively, FC, FCHT, and HT. For the VG and AS models, as stated previously, only the FC approximation will be considered here, as a consequence, the redundant FC suffix will be then omitted. However, it should be noted that present implementation is able to handle transparently FCHT and HT calculations with the VG and AS models. All such models can be combined with computations of one-photon electronic spectra induced by either electric or magnetic transition dipoles or by their mutual interaction. Table 1 provides an overview of the information needed as input for each of the considered models. It is noteworthy that all approaches require the optimized geometry for the initial state along with the calculation of its Hessian matrix. However, they differ significantly for the data required for the final state, whose PES can be built only from the ground of its energy gradient for the simplest VG model or from a full geometry optimization and harmonic analysis for the most demanding one. As a consequence, the choice of the model might have a large impact on the total computational times.

The simplest VG model requires only the energy gradient of the excited state to be calculated at the geometry of the ground state. In a time-dependent perspective, the VG approach is related to the effect of short-term dynamics on the spectra, so it is expected to reproduce well the low-resolution spectrum shape. On the other side, it does not account for the changes in vibrational frequencies and for the normal modes' mixing between the excited and ground electronic states. Because of its characteristics, the VG model provides the most up to date and feasible approach for the studies of the spectrum in a broad energy range and/or for macromolecules. At variance, the AS model requires the determination of the equilibrium structure for the final state but not the frequencies, so it might be considered as a solution for cases where the main interest is in the spectral features close to the transition origin, but no precise frequencies are required. It should be noted that both VG and AS models constrain the total zero-point vibrational energy to be the same in the initial and final states. At variance, they evaluate differently the transition energy between the minima of the initial and final states, which is more accurately computed within the adiabatic framework. However, in both cases, if ZPVE effects are introduced and excited-state frequencies are computed from second derivatives of the excited-state PES, sensible differences can be found, introducing shifts of the final energy levels that are often larger than 0.1 eV.

Finally, the AFC or AFCHT approaches are best suited when an accurate reproduction of the excited-state frequencies, a fine structure of the spectra, and a good estimate for absolute positions of vibronic bands are necessary. However they are rather expensive in terms of computational costs since they involve geometry optimization and Hessian computations in the excited electronic state. Vertical approaches based on a full calculation of the Hessian of the excited state can be convenient in some cases, but they have not been implemented yet and will be the subject of future work. It is worthwhile to note that, in the present implementation, if ab initio computations are performed at the TD-DFT level, then frequency calculations are performed by numerical differentiation of the analytical TD-DFT energy gradients, so the transition dipole electric and magnetic moments as well as their derivatives are directly available from the frequency calculations. However, if analytical second derivatives with respect to the geometrical parameters are available for the method used to compute the energy of the excited electronic state, then the numerical calculation of frequencies must be requested in order to use the Herzberg−Teller approximation of the transition dipole moment. The implementation of the presented computational tool into a general purpose computational package facilitates its use for nonspecialist. However, the presented approach is not limited to quantum mechanical treatments available in the Gaussian package and besides the internal ab initio data user-defined data, e.g. for the vibrational frequencies and/or energies of each electronic state as well as for the transition dipole moments can be provided as input, increasing the flexibility of the implementation. Such a feature is of particular relevance for hybrid approaches where geometries, Hessian computations, and related derivatives of transition dipole moments are obtained at a lower level of theory (e.g., DFT with a medium-size basis set), while energy of electronic transition and related transition dipole moments are further refined by more accurate computations (e.g., DFT with larger basis sets and RI-CC2). Moreover, a number of parameters are then available to fine-tune the spectra layout or to control the prescreening, among others. About the former, it is possible to not only define the spectrum bounds but also the distribution functions used to simulate the spectrum broadening. To this purpose, both Lorentzian, simulating a homogeneous broadening, and Gaussian functions for the inhomogeneous broadening, due to solvent or temperature effects, are available. Moreover, the quality of the spectra simulations is ruled by parameters set for the prescreening method, and various aspects of spectrum convergence with respect to these parameters have been discussed in detail in ref 1. The inclusion of temperature has led to additional functionalities. The two major parameters are the temperature and the minimum Boltzman population (MinPop). The second one sets the minimum population a vibrational state must have with respect to the Boltzmann population of the vibrational ground state (set to 1) to be taken into account as an origin of the transitions. As mentioned before, the initial vibrational states are divided by class and then grouped in sets. Each set is treated independently with a specific total intensity $I^{tot}$ calculated for it. The spectrum convergence as well as the assignment is given for each set. By default, only the final spectrum is printed. However, when needed, specific spectra for each class of the initial and/or final states are available.

In this work, we will mainly concentrate on newly introduced features with respect to the previous Gaussian implementation[1] and discuss the problems related to their applicability to quite different spectroscopic studies, as discussed further in Section 5. The applications presented here are confined to transitions between bound electronic states. Nonetheless, we highlight that the developed methodology and all the different implemented methods can be utilized also for the simulation of photoionization and photoelectron spectra, with the approximation of considering the transition probability independent of the free electron final state and kinetic energy. In this limit, with the current facilities within the Gaussian package, it is possible to obtain the FC envelope of the spectra in arbitrary units.

## 3. Computational Chemistry Models

The computational chemistry methods have been chosen according to the system under study in order to find a satisfactory balance between feasibility of calculations and accuracy of results. First, a small chiral molecule, *R*-methyloxirane (RMO), has been chosen to assess the computational models, which will be used in further studies, with particular reference to the accuracy of the computed rotatory strengths. For such a reason, the ground and excited electronic state computations have been carried out using DFT and TD-DFT,[23] respectively, with the standard B3LYP,[49] and its longe-range corrected extension CAM-B3LYP,[50,51] functionals in conjunction with N07D,[52,53] N07T,[53,54] and aug-cc-pVTZ basis sets.[55,56] Additionally, for the calculations with CAM-B3LYP, basis sets up to aug-cc-pV5Z have been considered, and their effect on the computed vertical excitation energies and on the rotatory strengths is discussed in the detail. For (*R*)-(+)-3 methylcyclopentanone (R3MCP), we have chosen to apply the CAM-B3LYP/aug-cc-pVDZ computational model, in line with the detailed studies of its ECD spectra by some of us.[5−7,57] Additionally, for R3MCP the convergence of the computed values of the transition dipole moments with respect to the basis set has been evaluated by the comparison of aug-cc-pVDZ results with the N07 double- and triple-ζ, augmented by one set of diffuse functions on the heavy atoms[53] (N07Ddiff and N07Tdiff). For chlorophyll *a*1, the ground-state structure and the harmonic frequencies have been computed with the B3LYP/N07D model, while the forces in first eight singlet excited electronic states have been computed at the TD-DFT level with the CAM-B3LYP[50] functional and the N07D basis set augmented by one set of diffuse functionals on heavy atoms (N07Ddiff), as recommended in excited studies of vinyl radical.[58,59] The vertical excitation energies have been computed as a difference between ground- and excited-state energies computed with CAM-B3LYP/TD-CAM-B3LYP, respectively, with a N07Ddiff basis set, while solvent effects

***Table 2.*** Vertical Electronic Excitations (VE in eV) and Rotatory Strengths (R in cgs Computed with the Length Gauge) Computed For RMO with the TD-CAM-B3LYP Density Functional and Basis Sets Ranging from N07D to aug-cc-pV5Z

| | | AV5Z | AVQZ | AVTZ | N07Tdiff | N07T | N07Ddiff | N07D |
|---|---|---|---|---|---|---|---|---|
| state | exp | | | | VE | | | |
| 2A | 7.07 | 7.14 | 7.15 | 7.15 | 7.13 | 7.15 | 7.15 | 7.69 |
| 3A | 7.7 | 7.46 | 7.48 | 7.49 | 7.46 | 7.49 | 7.53 | 8.04 |
| 4A | | 7.56 | 7.58 | 7.58 | 7.55 | 7.56 | 7.61 | 8.19 |
| 5A | | 7.71 | 7.73 | 7.74 | 7.72 | 7.72 | 7.78 | 8.58 |
| 6A | 8.5 | 7.85 | 7.86 | 7.86 | 7.87 | 7.89 | 7.89 | 8.67 |
| 7A | | 8.28 | 8.30 | 8.30 | 8.32 | 8.32 | 8.37 | 8.80 |
| 8A | | 8.32 | 8.34 | 8.36 | 8.36 | 8.38 | 8.45 | 8.95 |
| | exp | | | | R | | | |
| 2A | −12.56 | −15.35 | −16.19 | −16.63 | −16.23 | −17.01 | −19.89 | −3.02 |
| 3A[a] | 6. 98[a] | −7.40(7.76) | −7.08(9.06) | −6.81(9.73) | −6.46(9.71) | −6.51(9.08) | −3.53(16.13) | 3.58(4.85) |
| 4A[a] | | 8.47 | 8.64 | 8.72 | 8.89 | 8.75 | 10.12 | 10.66 |
| 5A[a] | | 6.69 | 7.50 | 7.81 | 7.27 | 6.84 | 9.54 | −9.39 |
| 6A | | 10.55 | 10.87 | 11.03 | 10.50 | 11.04 | 11.80 | 14.07 |
| 7A | | −1.50 | −0.11 | 0.86 | 0.04 | 0.72 | 0.53 | −5.86 |
| 8A | | −15.95 | −18.39 | −19.83 | −16.02 | −13.58 | −19.74 | −18.42 |

[a] Comparison with experiment is made by summing the transitions to the 3A−5A Rydberg states, shown in parentheses.

on the UV−vis spectrum of chlorophyll *a*1 have been introduced by a polarized continuum medium, as described by the conductor-like polarizable continuum model (CPCM)[60] model. For the OPA and ECD studies of (*Z*)-8-methoxy-4-cyclooctenone (MCO), only the lowest, $\pi^* \leftarrow n$ electronic transition has been considered, thus, the TD-B3LYP/B3LYP//N07Ddiff model, which is able to provide accurate results for the lowest excited states has been applied. All calculations have been performed with a locally modified version of the Gaussian suite of quantum chemistry programs.[61]

## 4. Validation of the DFT/N07 Model for Computations of ECD Spectra

It is widely recognized[62−65] that computations of ECD spectra are particularly challenging for the quantum mechanical treatments. Hence, the definition of a reliable and feasible computational tool, able to handle studies of ECD spectra, is of relevance to the general applicability of the presented model. Recently, some of us have presented a DFT/N07D model[52,66,67] which has been further validated for a broad range of computational spectroscopy studies: electron spin resonance (ESR), infrared (IR), UV−vis.[58,59] In this work, we have chosen to check the performance of the N07D basis set for the challenging case of ECD spectroscopy as well as the new triple-ζ basis set in the N07 family,[53,54] that is introduced here. In these benchmark studies, we will not only consider the absolute values of the rotatory strengths but also the accuracy of the computed derivatives of the magnetic and electric dipole moments necessary for the inclusion of HT effects into the simulated ECD spectra.

**4.1. Computation of Rotatory Strengths for *R*-methyloxirane.** RMO stands as a popular example, due to its small size, for the benchmark studies of properties of chiral systems.[63,64,68] In particular, the effect of the basis set on the vertical excitation energies and on the rotatory strengths computed at the B3LYP level with the aug-cc-pVXZ (X = D, T, Q) basis sets (shortly named from now on “AVXZ”) has been studied, pointing out the importance of diffuse functions.[63] The results gathered in Table 2 show that all basis sets starting from N07Ddiff predict the vertical excitation energies with a comparable accuracy (within 0.1 eV) and also agree fairy well with the experimental data. Such a finding further confirms the applicability of the N07Ddiff basis to the study of excited states, in line with recent investigations.[58,59] However, as already discussed, computations of rotatory strength are much more demanding with respect to the basis set convergence. It should be noted that none of the theoretical results presented in Table 2, match precisely the experimental values, but such disagreement can be attributed to solvent and vibrational effects as already postulated in previous studies.[63,68] For this reason, we will consider in the following discussion, the results from CAM-B3LYP/aug-cc-pV5Z computations as reference data. It can be observed that the standard N07D basis set is not even sufficient for qualitative studies on the rotatory strength. However, a significant improvement is obtained by adding a set of diffuse functions on heavy atoms, which leads to a semiquantitative agreement, at the expense of slightly more demanding computations. Results can be further improved with the N07T basis set, in particular its augmented version. In fact, the N07Tdiff basis set provides rotatory strengths of accuracy comparable to the much more computationally demanding AVQZ basis set (190 vs 596 basis functions, respectively), and on the whole, closer to the most expensive basis set considered, aug-cc-pV5Z (988 basis functions), than the triple-ζ basis set from the aug-cc-pVXZ family (AVTZ, 322 basis functions). Additionally, we have chosen to test the performance of the standard B3LYP functional for rotational strength computations and to compare the results to the ones obtained with the CAM-B3LYP functional, as the latter provides rotatory strengths in good agreement with experiment, in particular, if sufficiently large basis sets are applied. Results gathered in Table 3 show that, as expected, the B3LYP provides qualitatively correct values only for lower lying excited states of valence character. In conclusion, the results discussed above show that the CAM-B3LYP/N07 model is able to provide qualitatively correct results for all excited electronic states starting with the augmented basis set of double-ζ quality (N07Ddiff), while further refinements

***Table 3.*** Vertical Electronic Excitations (VE in eV) and Rotatory Strengths (R in cgs) Computed for RMO with the TD-CAM-B3LYP and TD-B3LYP Density Functionals and aug-cc-pVTZ and N07Tdiff Basis Sets

| | | TD-CAMB3LYP | | TD-B3LYP | |
|---|---|---|---|---|---|
| | | N07Tdiff | AVTZ | N07Tdiff | AVTZ |
| state | exp | | VE | | |
| 2A | 7.07 | 7.13 | 7.15 | 6.53 | 6.56 |
| 3A | 7.7 | 7.46 | 7.49 | 6.96 | 6.99 |
| 4A | | 7.55 | 7.58 | 7.00 | 7.03 |
| 5A | | 7.72 | 7.74 | 7.08 | 7.11 |
| 6A | 8.5 | 7.87 | 7.86 | 7.37 | 7.36 |
| 7A | | 8.32 | 8.30 | 7.72 | 7.76 |
| 8A | | 8.36 | 8.36 | 7.82 | 7.82 |
| | | | R | | |
| 2A | −12.56 | −16.23 | −16.63 | −18.76 | −19.37 |
| 3A[a] | 6. 98[a] | −6.46(9.71) | −6.81(9.73) | 12.48(15.89) | 13.01(16.43) |
| 4A[a] | | 8.89 | 8.72 | 1.20 | 0.88 |
| 5A[a] | | 7.27 | 7.81 | 2.20 | 2.55 |
| 6A | | 10.5 | 11.03 | 8.40 | 8.80 |
| 7A | | 0.04 | 0.86 | 2.67 | 2.74 |
| 8A | | −16.02 | −19.83 | −13.55 | −3.50 |

[a] Comparison with experiment is made by summing the transitions to the 3A−5A Rydberg states, shown in parentheses.

can be introduced by the N07Tdiff one. Additionally, it should be stressed that the B3LYP functional can also be applied to cases where only transitions to the lowest lying electronic states of valence character are considered.

**4.2. Accuracy of the Derivatives of the Transition Dipole Moments Computed at The DFT/N07 Levels: Line Shape Convergence of ECD−AFCHT Spectrum for the $S_2 \leftarrow S_0$ Electronic Transition of ax-R3MCP.** In the present approach, it is possible to improve the quality of the OPA, OPE, and ECD spectra computed within the AFC approximation by considering changes of the transition dipole moments with the geometry. It has been already mentioned that, as far as the computations of excited electronic state frequencies are performed at the TD-DFT level, the inclusion of the Herzberg−Teller term does not require any additional quantum mechanical calculations with respect to the AFC. For such reason, it should be recommended to perform also FCHT computations whenever possible, in particular, for weakly allowed electronic transitions or for cases when ECD spectra have to be studied. For the latter, it has already been demonstrated that inclusion of the HT term might change the sign of some vibronic lines, as for example, the case of axial-methyl conformer of (*R*)-(+)-3 methylcyclopentanone (ax-R3MCP).[6,57] The higher sensitivity of ECD computations to the transition dipole moment approximation, compared to OPA and OPE, is related to the dot product of two different transition dipole moments and, more precisely, to their relative orientation, which stands as an additional factor. For instance, when the mutual orientation of electric and magnetic dipole moments is close to 90 degrees, small changes might introduce a sign reversal of the computed rotatory strength. For this reason, it is important to check not only how the basis set influences the computed rotatory strength at the equilibrium but also more subtle effects on the accuracy of the computed derivatives of transition dipole moments, which might cause significant changes of the spectral lines. The analysis of the basis set effects on the rotatory strength, presented in Section 4.1, revealed that fairly accurate results are provided by computations with the CAM-B3LYP functional in conjunction with the N07Tdiff basis set. This basis set is small enough to allow the simulation of ECD spectra for medium-size systems with adiabatic approaches, requiring excited electronic state frequency computations. However, having in mind larger systems, it is important to check the performance of the smaller N07Ddiff basis set and to compare these results with CAM-B3LYP/aug-cc-pVDZ computations.[57] When studying the ECD spectrum related to the $S_2 \leftarrow S_0$ electronic transition of the ax-RMCP, it appeared that some of the vibronic transitions changed sign by including the HT term.[6,57] As a consequence, we have chosen ax-RMCP for the benchmark study of the basis set effect on the computed transition dipole moment derivatives and on their influence of the spectra line shape. Panel a in Figure 1 shows the ECD spectra in a 0−4000 $cm^{-1}$ energy range from the transition origin, computed with the three different basis sets under study with the AFC and AFCHT approximations and convoluted with Lorentzian functions with full width at half-maximum (fwhm) of 0.005 eV (in line with previous studies). It is immediately visible that several bands change sign when the HT term is taken into account, and such an effect is consistently obtained for the main spectrum features by all basis sets. Looking more into detail, as shown in panel b of Figure 1 and in the stick spectrum presented in Figure 2, it is possible to find some minor vibronic contributions which differ in sign depending on the basis set. The most visible difference is related to the $\langle 0|2^1\rangle$ transition, where computations at the FCHT level with the aug-cc-pVDZ basis set predicts a negative vibronic band in line with FC approximation, while for the N07 basis sets, its contribution to the spectra is canceled through the HT term. However, despite minor discrepancies, we shall conclude that all studied basis sets provide comparable results for the ECD-FCHT spectra. In particular, the good qualitative agreement obtained with the N07Ddiff basis set justifies its applicability to the studies of vibrationally resolved ECD spectra within the FCHT approximation, which requires expensive computations of the vibrational frequencies in the excited electronic states. It should also be stressed that the hybrid approach with transition dipole moment derivatives computed with basis

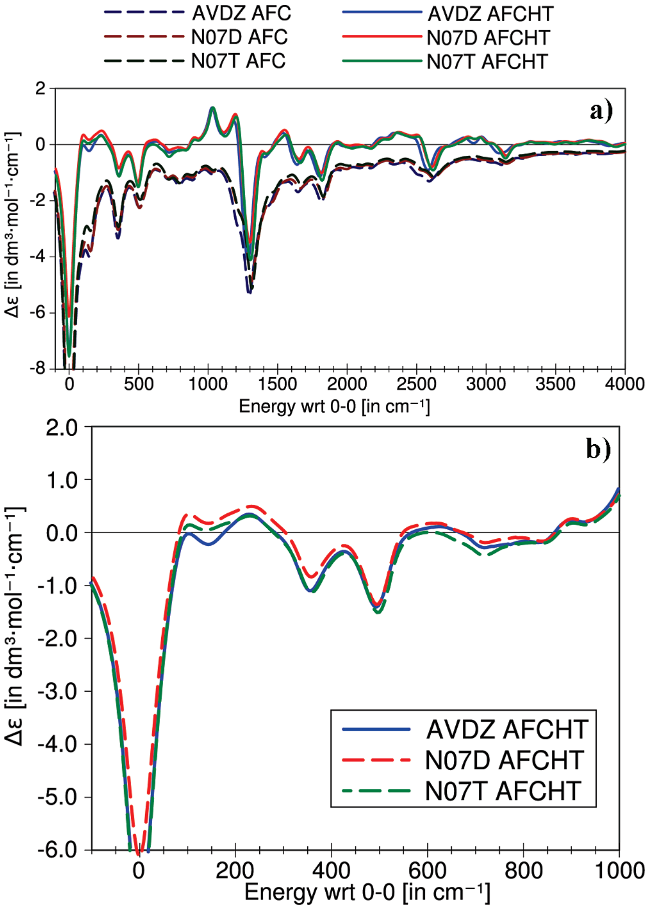


***Figure 1.*** Convergence of the ECD spectrum line shape with respect to the basis set. Accuracy of the derivatives of the transition dipole moments computed at the TD-CAM-B3LYP level with aug-cc-pVDZ (AVDZ), N07Ddiff (N07D), and N07Tdiff (N07T) basis sets. Adiabatic Franck−Condon (AFC) and Franck−Condon Herzberg−Teller approaches (AFCHT) for the $S_2 \leftarrow S_0$ electronic transition of ax-R3MCP (color online). The bands are convoluted with Lorentzian functions of 0.05 eV fwhm and with the spectra span energy range of 0−4000 cm$^{-1}$ (panel a) or 0−1000 cm$^{-1}$ (panel b) with respect to the 0-0 transition.

sets of N07D quality, combined with equilibrium properties evaluated at higher level of theory, can offer a noteworthy refinement for larger systems. The current implementation allows such a hybrid scheme through reading of appropriately defined transition dipole moments and their derivatives from the input stream. However, it should be remembered that on the ground of the perturbative HT theory of intensity borrowing,[35] such hybrid approach is not expected to be reliable when the energy gap between the lending and borrowing states change considerably with the basis set.

## 5. Simulated One-Photon Electronic Spectra

The integrated approach to compute vibrationally resolved one-photon electronic spectra can be applied to a large variety of systems ranging from small molecules in the gas phase to macrosystems in condensed phases, whenever the nonadiabatic couplings are negligible and the harmonic approximation is reliable. In this section, the $S_2 \leftarrow S_0$ one-photon ECD spectra of R3MCP are used to present various aspects of the convergence of the simulated results (Sections 5.1 and 5.2), while the last two subsections: the UV−vis absorption spectrum in the 250−700 nm range of chlorophyll *a*1 (Section 5.3) and the ECD spectrum of the $\pi^* \leftarrow n$ transition of MCO (Section 5.4) are presented in order to illustrate the flexibility of the integrated approach and its applicability to larger systems.

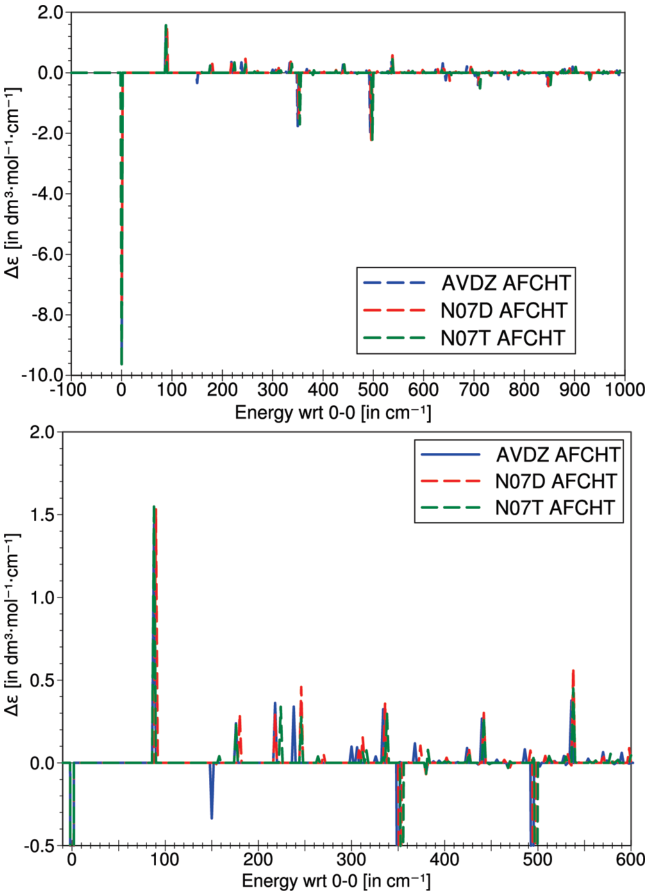


***Figure 2.*** Convergence of the ECD spectrum line shape with respect to the basis set. Accuracy of the derivatives of the transition dipole moments computed at the TD-CAM-B3LYP level with AVDZ, N07D, and N07T basis sets. The stick spectra of the AFCHT for the $S_2 \leftarrow S_0$ electronic transition of ax-R3MCP (color online) span an energy range of 0−1000 cm$^{-1}$ (panel a) or 0−600 cm$^{-1}$ (panel b).

**5.1. Vertical and Adiabatic Approaches to Compute Electronic Spectra: Case of OPA, OPE, and ECD Spectra for ax-R3MCP.** As already shown in previous sections and discussed in refs 6 and 57, the $S_2 \leftarrow S_0$ electronic transition of the ax-RMCP is an interesting case where HT effects can lead to a change of the sign with respect to the simple FC approach, in case of ECD spectra, but have a lower impact in OPA simulations. Figures 3, 4, and 5 show the simulated OPA, OPE, and ECD spectra obtained with various approximations related to changes between the electronic states within the electronic transition. In fact, for OPA spectra, it is immediately visible that the simplest VG approach yields spectrum line shapes in qualitative agreement with more demanding computations performed within the adiabatic framework, so it can be sufficient to reproduce correctly the general features of the experimental spectra. However, as clearly shown in Figure 4, such an agreement

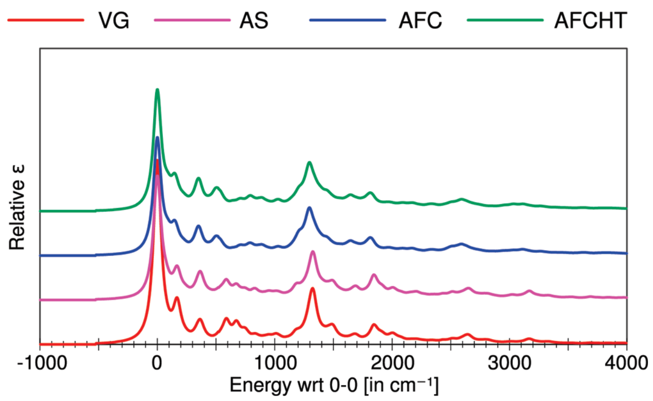


***Figure 3.*** Convoluted OPA spectra of the $S_2 \leftarrow S_0$ electronic transition of ax-R3MCP (color online) computed with vertical and adiabatic approaches: VG, AS, AFC, and AFCHT.

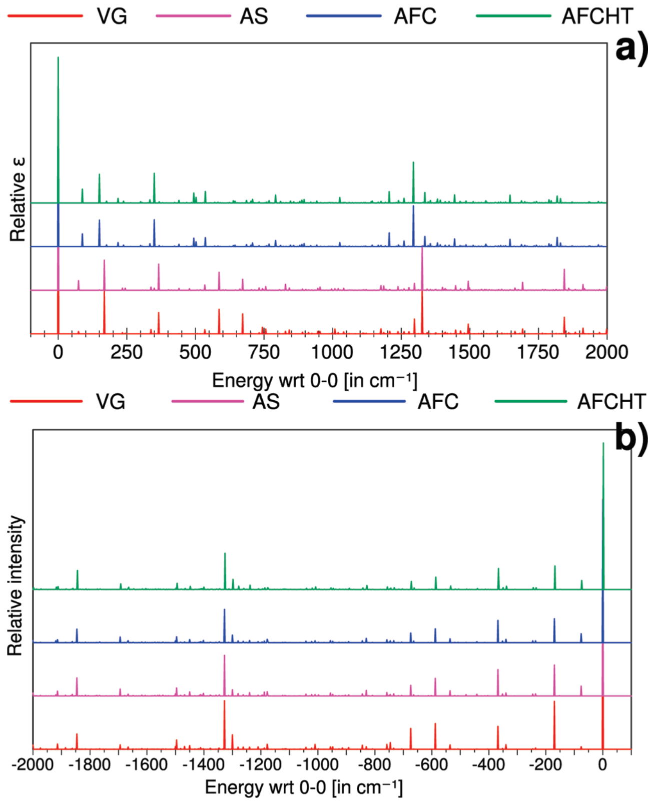


***Figure 4.*** Stick OPA and OPE (panel a and b, respectively) spectra of the $S_2 \leftarrow S_0$ electronic transition of ax-R3MCP (color online) computed with vertical and adiabatic approaches: VG, AS, AFC, and AFCHT.

cannot be expected if the fine structure is also needed. Moreover, larger deviations can be expected for molecules (and/or states) showing larger changes in frequencies and/or significant Duschinsky mixings upon excitation. As an example, in ref 57, sensible differences between VG and AFC results have been documented for the $n \rightarrow \pi^*$ ($S_1 \leftarrow S_0$) electronic transition of both ax- and eq- conformers of RMCP. Such a transition involves the lone pair of the oxygen and the $\pi^*$ orbital residing on the CO bond, so that the CO stretch is responsible for the most prominent progression of the spectrum, and its vibrational frequency strongly decreases upon excitation. In this situation, models like VG that do not take into account this effect predict too-large spacings between the bands, while application of the AFC model results in a significant improvement of the agreement with the experiments. The situation is different in the case of the OPE spectrum, since the fine structure corresponds to the vibrational frequencies in the ground state, which are correctly taken into account in all models. In fact, panel b of Figure 4 shows that VG, AS, FC, and FCHT predict the same positions for the vibrational transitions. However the general pattern is different, as some bands are missing in VG or varying in intensity.

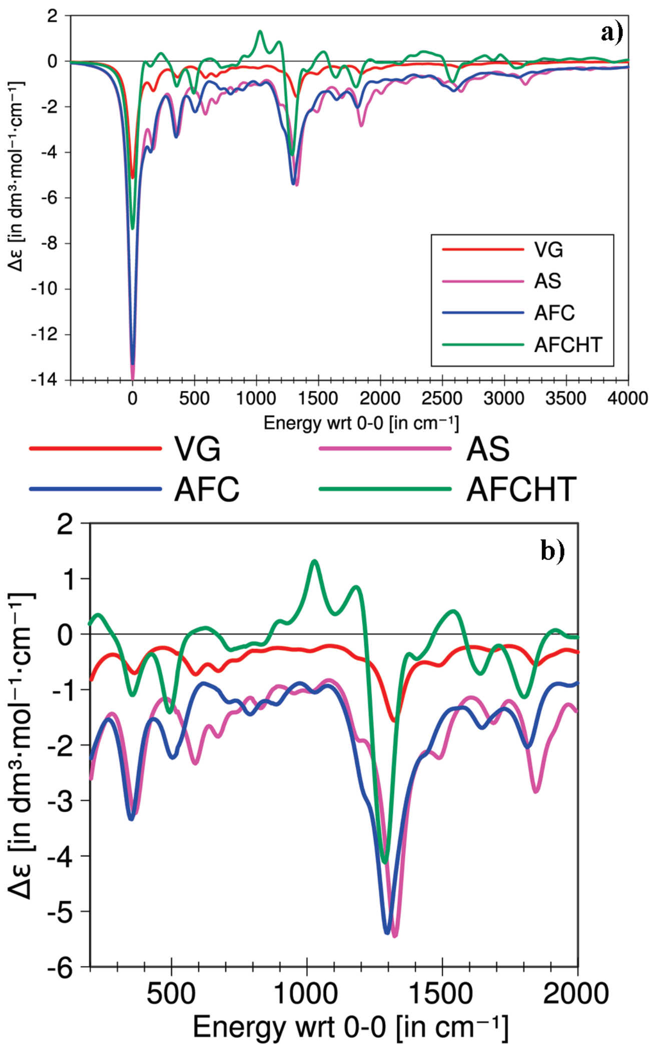


***Figure 5.*** ECD spectra of the $S_2 \leftarrow S_0$ electronic transition of ax-R3MCP computed with vertical and adiabatic approaches: VG, AS, AFC and AFCHT (color online), which span an energy range of 0−4000 cm$^{-1}$ (panel a) or 200−2000 cm$^{-1}$ (panel b) with respect to the 0-0 transition.

However, the situation is more complex for ECD spectra; in this case, much more pronounced differences between VG and AS or AFC computations can be observed. In fact, even the AFC and AFCHT spectra line shapes differ significantly when the HT contributions change the sign of some of vibronic transitions, an effect which obviously cannot be reproduced by any FC based approaches: VG, AS, or AFC. Such findings clearly show the higher sensitivity of ECD to the approximation of the transition dipole moment used to compute the spectra. Thus, while VG approaches can be also applied to ECD, the VG-ECD results need to be considered qualitative at most. It must be noticed that, in ref 5, it has been shown that the VG method with the inclusion of HT effects can reproduce correctly the change in the sign of vibronic transitions, even if differences are visible in the details of the spectrum. Nonetheless, we highlight once more that, at the state-of-the-art level, the computational effort to compute the derivatives of the transition dipole moment is the same as the one required to obtain all the necessary data to compute the spectrum at the AFCHT level, i.e., by fully including all the possible spectral features within the harmonic approximation.

**5.2. Temperature Effects and Convergence of Line Shapes for One-Photon Electronic Spectra Simulated in Ambient Conditions. Case of OPA and ECD Spectra for ax-R3MCP.** In the current implementation, it becomes possible to include the temperature effects on the spectrum shape, a feature particularly important for the comparison with experimental data performed at ambient conditions. In general, when temperature-dependent spectra are to be simulated by a time-independent approach, like the one pursued here, additional issues arise in choosing the set of vibrational states from which electronic transitions will take place. This choice is defined by the probability of the molecule to be in each initial vibrational state given by its Boltzmann population. Then, a suitable threshold for the minimum population of the vibrational states to be included into the sets considered in computations needs to be chosen. This threshold is defined with respect to the Boltzmann population of the vibrational ground state (set to 1) and can be modified freely (through the MinPop keyword). The intensity and line-shape convergence, with respect to the population threshold, will also be discussed on the examples of OPA−AFC and ECD−AFCHT spectra for the $S_2 \leftarrow S_0$ electronic transition of the ax-RMCP, simulated at 298 K, which are shown in Figures 6 and 7, respectively. It can be immediately observed that temperature has a similar effect in both cases and that the increasing number of initial vibrational states (lower percentage of ground-state Boltzmann population required for state to be included into set) modifies mainly the part of the spectra close to the 0-0 transition, while in the higher energy wing, spectra do not differ significantly from the one computed at 0 K. Additionally, we can note that some temperature effects are already visible even if a limited number of vibrational states is taken into account (MinPop = 25%) and that increasing the number of initial vibrational states leads to a decrease in the intensity of the main transitions, coupled with an increase in intensity for the less pronounced features. Nevertheless, for the AFC approximation, the total spectrum intensity remains constant at the value corresponding to the simulations at 0 K. However, usually the spectrum shape is the property of interest and, as shown in Figure 6, the spectra computed with a minimum population set of 25 and 10% are already quite similar when the intensity with respect to the total Boltzmann population is considered. The latter finding is also valid for ECD spectra, as shown in Figure 7, for which the line-shape convergence does not require the inclusion of a very high number of vibrational initial states, which would, in turn, increase steeply the computational cost. In line with this finding, we have chosen to set to 10% the default value of the Boltzmann population to be considered, however, such a value might be freely modified whenever needed. Nevertheless, it should be noted that in some cases, where normal modes are significantly displaced or noteworthy mixing between them is observed as well as in cases where strong

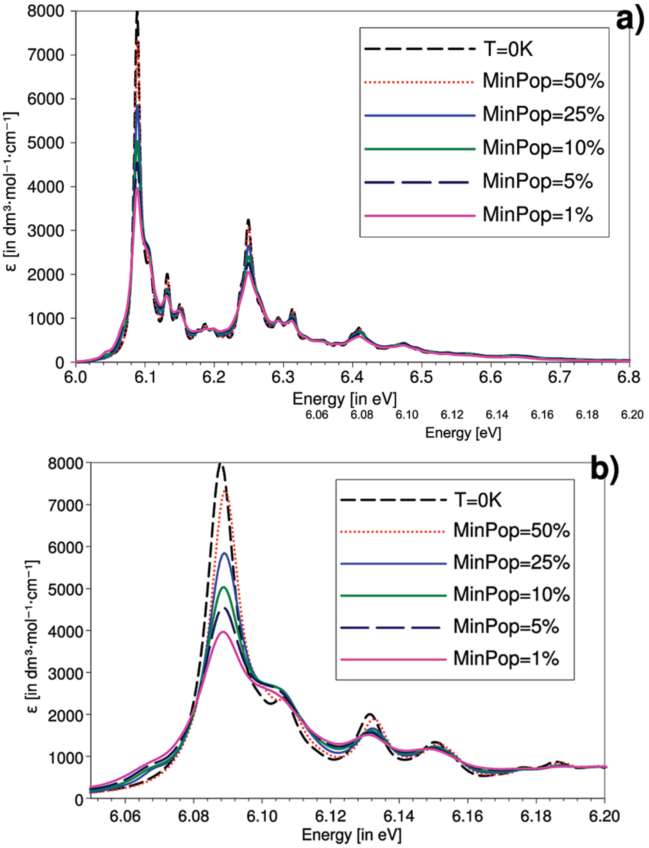


***Figure 6.*** Temperature effects and convergence of the line shapes for one-photon electronic spectra simulated in ambient conditions (298 K) are shown with the case of the OPA spectra of ax-R3MCP (color online), simulated within the AFC framework. The spectra span an energy range of 6.0−6.8 eV (panel a) or 6.06−6.20 eV (panel b).

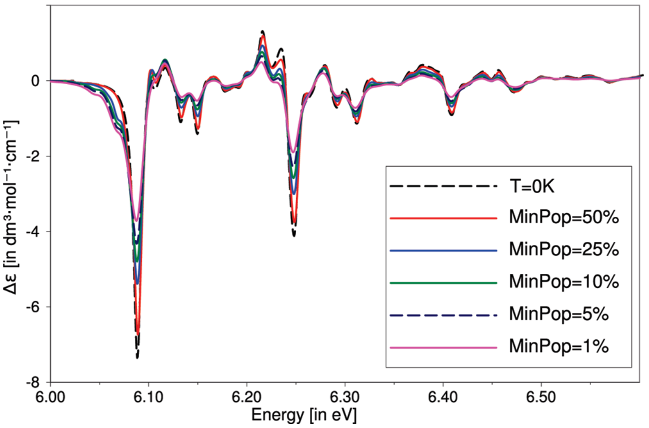


***Figure 7.*** Temperature effects and convergence of the line shapes for one-photon electronic spectra simulated in ambient conditions (298 K) are shown with the case of the ECD spectra of ax-R3MCP (color online), simulated within the AFCHT framework.

progressions are induced by the hot vibrational modes, such a threshold might not be satisfactory.

**5.3. UV−vis Spectrum of Chlorophyll *a*1.** The understanding of the molecular mechanism of light harvesting in the photosystem II is one of the subjects intensively studied both experimentally and theoretically. For the latter, recent developments within the computational approaches and within the increased computational resources allow, at present, studies at the QM level in ground as well as in excited states. In such a way, a new level of accuracy has become available, and it may be expected that QM computations of optical properties combined with spectroscopic experiments will contribute to shed further light on this phenomenon.[69] In the present work, we have chosen to study the UV−vis spectrum of chlorophyll *a*,[70,71] which has been modeled in the current approach by chlorophyll *a*1, a large molecule with 46 atoms and 132 normal modes. For such a system, fully QM simulations of vibronic spectra within the AFC or AFCHT frameworks are already possible but still computationally demanding, in particular, if large energy windows, encompassing several electronic transitions, need to be studied. This situation can be significantly improved with the VG approach where, to simulate vibrational effects on the spectrum line shape only computations of excited-state forces are required, allowing a relatively cheap and straightforward computation of low-resolution electronic spectra for large molecules in gas phase and in solution. Here, we present such a study for chlorophyll *a*1, for which electronic QM computations have been performed at the DFT/N07D level, and the effect of the methanol solvent has been included by means of the polarizable continuum model, where the solvent is represented by a homogeneous dielectric polarized by the solute and placed within a cavity built as an envelope of spheres centered on the solute atoms.[60] The solvent has been described in the nonequilibrium limit where only its fast (electronic) degrees of freedom are equilibrated with the excited-state charge density, while the slow (nuclear) degrees of freedom remain equilibrated with the ground state. This assumption is well adapted to describe the broad features of the absorption spectrum in solution due to the different time scales of the electronic and nuclear response components of the solvent reaction field.[3]

The simulated UV−vis spectrum in a 250−700 nm range has been obtained by summation of the contributions from transitions to the first eight singlet excited electronic states. It can be noted that the new features of the integrated procedure to compute electronic spectra, which reports results in the absolute values (see Section 2.1) instead of arbitrarily normalized intensity units, allow a more straightforward comparison of relative intensities of vibronic contributions from transitions to different electronic states. Figure 8 shows spectra simulated in a gas phase and a methanol solvent for chlorophyll *a*1, which are compared to the experimental data from the solution.[70,71] It is immediately visible that, while both computed spectra reproduce qualitatively the line shape of their experimental counterpart, a much better agreement, in particular for the absolute positions of vibronic bands, has been obtained for the one simulated in methanol. For the spectrum simulated in methanol solvent, a small 500 $\text{cm}^{-1}$ shift on the energy scale leads to a very good agreement with experiment, and such shifted spectrum is also depicted in Figure 8. For the gas-phase spectrum, the application of a uniform energy shift of 1500 $\text{cm}^{-1}$ allows a good match in an energy region around 650 nm, but in this case, the position of most pronounced band is still blue-shifted with respect to the experiment. At this point, we would like to stress that accurate prediction of electronic energies is still a very difficult task even for the most advanced computational approaches, while TD-DFT/DFT//N07D computations have already proven to provide very reasonable estimates of the relative energetics of the electronic states.[58,59] The current study shows also the importance of the direct inclusion of solvent effects, which significantly improve the agreement with experimental data. Additionally, the simulated spectrum can be easily dissected into the single electronic transitions, as shown in Figure 9, allowing to analyze their individual contributions to the spectrum line

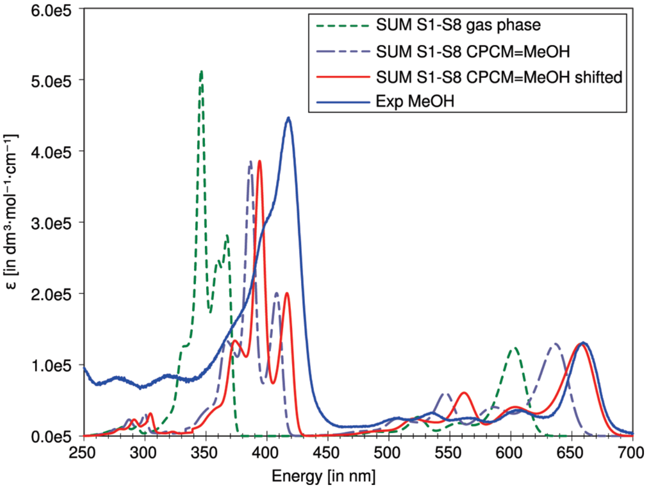


***Figure 8.*** The absorption spectrum of chlorophyll *a*1 in a 250−700 nm energy range, resulting from the sum of the transitions to the eight first singlet electronic states, is simulated in a gas phase and a methanol solution and compared to experimental data obtained in a methanol solvent.[70,71]

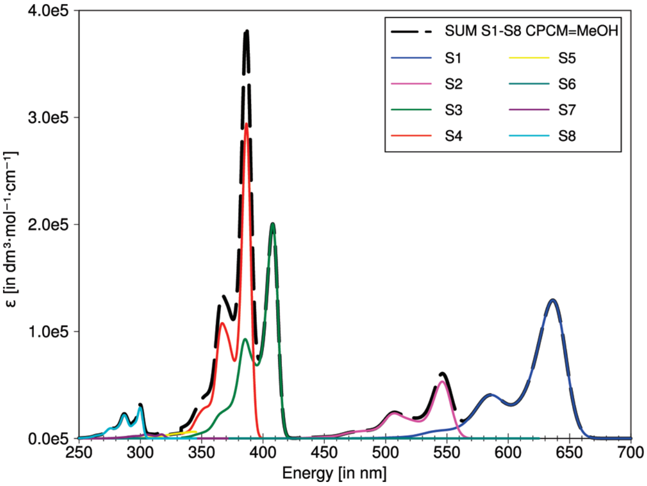


***Figure 9.*** The absorption spectrum of chlorophyll *a*1 (color online) in MeOH (CPCM) in a 250−700 nm energy range is dissected into the contributions of the single transitions.

shape. In fact, it can be immediately observed that in case of chlorophyll *a*1, the spectrum line shape is dominated by the contributions from transitions to the S1, S3, and S4 excited electronic states, with the non-negligible contributions from transitions to S2 and S8. In line with the general objective of the present work, we will mainly stress the ease and the feasibility of the presented integrated procedure, with results of very good quality obtained by the simple VG approach combined with relatively inexpensive QM studies available through the recently introduced DFT/N07D model. More detailed studies of the spectroscopic properties of chlorophyll *a*1 and of its radical cations are to be reported in a separate work.[72]

**5.4. Electronic OPA and ECD Spectrum of (*Z*)-8-Methoxy-4-cyclooctenone.** An unusual vibronic pattern in the ECD spectrum has been recently observed for the $\pi^* \leftarrow n$ electronic transition of the enantiopure (*Z*)-8-methoxy-4-cyclooctenone (MCO).[73] In the ECD spectrum, several bands of opposite sign have been observed, at variance with the single broad band found in the absorption spectrum. To analyze this issue, studies in the UV−vis energy range have been performed, aided by IR and VCD experiments and by computations at the DFT level. For the latter, only the vertical transition between the electronic states has been taken into account, and both absorption and ECD spectra have been simulated by applying a phenomenological broadening to the electronic stick data. Several conformers of MCO have been studied theoretically in order to obtain the best match with experimental spectra. On the basis of DFT computations for the ground state performed at the B3LYP/6-311+G(2d,p) level, the simulated IR and VCD spectra have been compared with their experimental counterparts, and the two most stable conformers were identified. However, the contribution from two additional, less stable conformers have been necessary to satisfactorily reproduce the unusual pattern of ECD spectra with the chosen model. It should be noted that our computations performed at the TD-B3LYP/B3LYP//N07diff level agree well with the results presented by Tanaka et al.[73] for the structure and energies of eight MCO conformers. In particular, conformers 1A and 1B are the most stable, and their energies differ only negligibly (within 0.2 kcal/mol), while all other conformers are less stable by at least 2 kcal/mol. Indeed, conformers 1A and 1B show an almost equal stability, so they should be observed in IR and VCD spectra as suggested. Here, we present the OPA and ECD spectra of MCO, which have been simulated considering only one of the two most stable conformers, labeled 1A in ref 71. The simulated absorption and the ECD spectra computed within the AFC and AFCHT frameworks are compared to their experimental counterparts[73] in Figures 10 and 11, respectively. The OPA spectra, depicted in Figure 10, show a single broad band both at the AFC and AFCHT level, in agreement with the experimental findings, and the inclusion of the HT contributions blueshift the maximum of the computed spectrum improving the agreement with the experiment. Obviously, the same result has been obtained when the AFC approximation has been used for the ECD spectroscopy and, indeed, ECD−AFC is a mirror reflection (due to negative sign of the rotatory strength) of OPA−AFC. At variance, the ECD spectrum computed within the AFCHT framework shows both positive and negative vibronic contributions, in line with experiment. Figure 11 compares the simulated ECD−AFCHT spectrum convoluted with a fwhm of 500 cm$^{-1}$ with the experimental

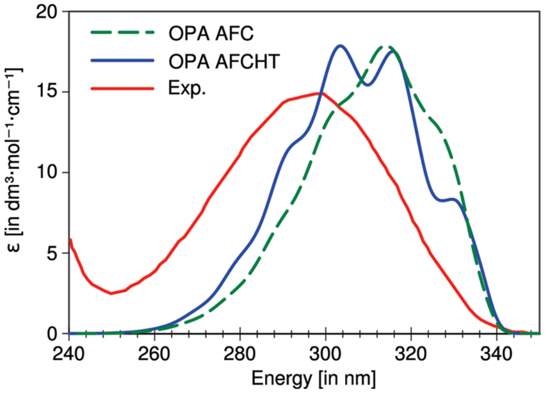


***Figure 10.*** Simulated (at AFC and AFCHT levels) and experimental absorption spectra of the $\pi^* \leftarrow n$ electronic transition of MCO. The theoretical spectra were convoluted with a fwhm of 500 cm$^{-1}$.

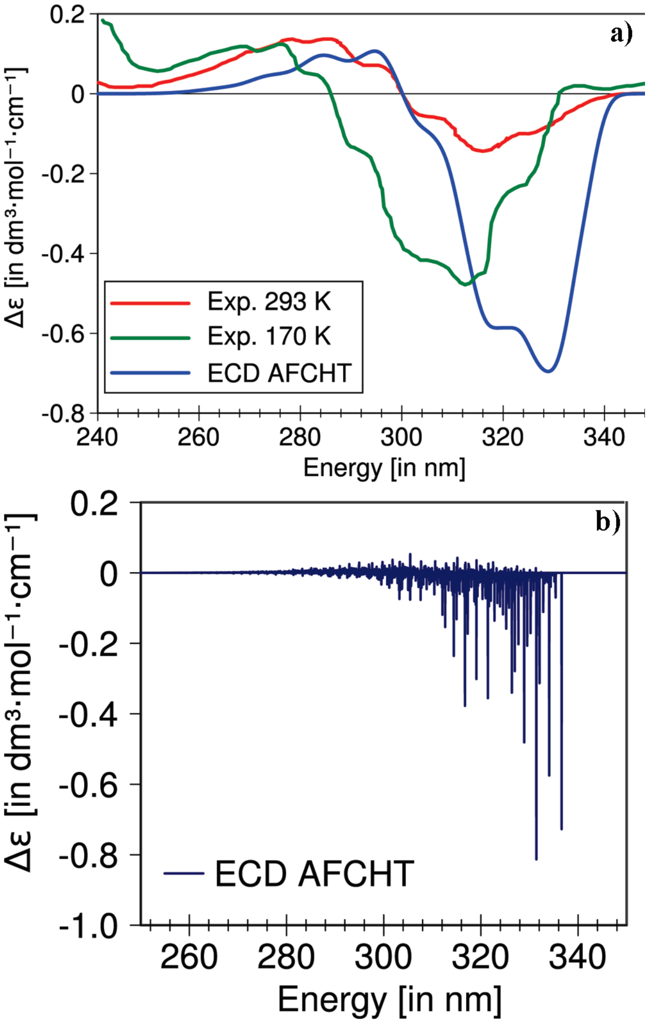


***Figure 11.*** Simulated (at the AFCHT level) and experimental ECD spectra of the $\pi^* \leftarrow n$ electronic transition of MCO. The theoretical spectra were convoluted with a fwhm of 500 cm$^{-1}$ .

data obtained at room temperature or by cooling the system to 170 K. In panel b of Figure 11, the theoretical stick spectrum is also shown, clearly indicating the presence of positive vibronic contributions. It is noteworthy that the spectrum simulated at 0 K is more similar to its experimental counterpart registered at 170 K than to the one taken at room temperature. More detailed insights of the temperature effects on the OPA and ECD spectra of MCO would require further investigations, which are beyond the scope of the present work. As mentioned above, IR and VCD studies confirmed the presence of another stable structure (1B) at experimental conditions, for which our computations yielded a very small FC integral between the vibrational ground sates. Nevertheless, even if the main spectral features are related to conformer 1A, it should not be excluded that contributions from 1B might slightly modulate the spectrum shape and that taking it into account could further improve the accuracy of the theoretical spectrum, with respect to the experiment. In any case, it should be noted that, within the simplified model where the vibrational effects are neglected,[73] as many as four conformers have been necessary for a qualitative reproduction of the experimental spectra. A change of sign can be sometimes observed in the energy window encompassed by a single electronic state[74] and is often attributed to the contribution of different conformers as in ref 73. However, the issues of possible changes in the sign of some vibronic lines of the ECD spectra due to the HT term have already been demonstrated in refs 6 and 57 and discussed in Section 4.2. Our results suggest that this is also the case for MCO, where the HT contribution indeed influences significantly the spectrum line shape and is necessary for a qualitative agreement with experimental data. Indeed, the simulation of the ECD spectrum within the FCHT framework can describe all its unusual features considering only the most stable isomer, at variance with the previous studies based on simple ECD spectra, simulated by applying a phenomenological broadening to the electronic stick data.

## Conclusions

A general approach for the simulation of vibrationally resolved one-photon electronic spectra has been implemented and applied to a variety of molecular systems, showing the high flexibility of the developed computational tool. The integration of all procedures within the same computational package allows for the fully automatic computation of vibrationally resolved electronic spectra. Despite the fact that our computational scheme has been tailored for large systems, it can be utilized as well to generate high-quality spectra for small systems, when nonadiabatic and anharmonic couplings are negligible, since it allows different levels of approximation for the computation of FC integrals. It should be noted that, even when nonadiabatic couplings can be neglected, several issues which are of general importance in many cases remain still open, like problems related to the presence of double-well potentials, large molecular displacements, or multimode couplings. However, all these issues are, in more general terms, related to the anharmonicity, and as we already stated, the simulations of vibrationally resolved electronic spectra with anharmonic models appropriately tailored for vibronic transitions are under active development. Notwithstanding the above limitations, we point out that, in the present work, we introduce an easy-to-use, general, and robust computational tool able to simulate good-quality spectra even for large systems with hundreds of normal modes, whenever harmonic approximation is reliable, paving the route to spectroscopic studies of systems of direct biological and/or technological interest, improving their interpretation and understanding.

**Acknowledgment.** This work was supported by Italian MIUR (PRIN 2006) and CNR (PROMO 2006). The large scale computer facilities of the VILLAGE network (http://village.unina.it) are kindly acknowledged.

**Supporting Information Available:** The generalized calculation of the spectrum convergence ($I^{\mathrm{calc}}/I^{\mathrm{tot}}$) for any couple of dipoles $\boldsymbol{d}_{\mathrm{A}}$ and $\boldsymbol{d}_{\mathrm{B}}$. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Barone, V.; Bloino, J.; Biczysko, M.; Santoro, F. *J. Chem. Theory Comput.* **2009**, *5*, 540–554.

(2) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Parandekar, P. V.; Mayhall, N. J.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fo, D. J. *Gaussian 09, Revision A.2*, Gaussian Inc.: Wallingford, CT, 2009.

(3) Santoro, F.; Improta, R.; Lami, A.; Bloino, J.; Barone, V. *J. Chem. Phys.* **2007**, *126*, 084509/1−13.

(4) Santoro, F.; Lami, A.; Improta, R.; Bloino, J.; Barone, V. *J. Chem. Phys.* **2008**, *128*, 224311/1−17.

(5) Lin, N.; Luo, Y.; Santoro, F.; Zhao, X.; Rizzo, A. *Chem. Phys. Lett.* **2008**, *464*, 144–149.

(6) Santoro, F.; Barone, V. *Int. J. Quantum Chem.* **2010**, *110*, 476–486.

(7) Lin, N.; Santoro, F.; Rizzo, A.; Luo, Y.; Zhao, X.; Barone, V. *J. Phys. Chem. A* **2009**, *113*, 4198–4207.

(8) Bloino, J.; Biczysko, M.; Crescenzi, O.; Barone, V. *J. Chem. Phys.* **2008**, *128*, 244105/1−15.

(9) de Groot, M.; Buma, W. J. *Chem. Phys. Lett.* **2007**, *435*, 224–229.

(10) de Groot, M.; Buma, W. J. *Chem. Phys. Lett.* **2006**, *420*, 459–464.

(11) Dierksen, M.; Grimme, S. *J. Chem. Phys.* **2004**, *120*, 3544/1−11.

(12) Pugliesi, I.; Tonge, N. M.; Hornsby, K. E.; Cockett, M. C. R.; Watkins, M. J. *Phys. Chem. Chem. Phys.* **2007**, *9*, 5436–5445.

(13) Tonge, N. M.; MacMahon, E. C.; Pugliesi, I.; Cockett, M. C. R. *J. Chem. Phys.* **2007**, *126*, 154319/1−11.

(14) Hazra, A.; Chang, H. H.; Nooijen, M. *J. Chem. Phys.* **2004**, *121*, 2125/1−12.

(15) Köppel, H.; Domcke, W.; Cederbaum, L. S. *Adv. Chem. Phys.* **1984**, *57*, 59–246.

(16) Köppel, H.; Domcke, W.; Cederbaum, L. The Multi-mode vibronic-coupling approach In *Conical Intersections, Electronic Structure, Dynamics and Spectroscopy*; World Scientific Publishing Co.: Singapore, 2004; pp 323−368.

(17) Nooijen, M. *Int. J. Quantum Chem.* **2006**, *106*, 2489–2510.

(18) Beck, H.; Jäckle, A.; Worth, G.; Meyer, H.-D. *Phys. Rep.* **2000**, *324*, 1–105.

(19) Christopher, P. S.; Shapiro, M.; Brumer, P. *J. Chem. Phys.* **2006**, *124*, 184107/1−11.

(20) Macak, P.; Luo, Y.; Ågren, H. *Chem. Phys. Lett.* **2000**, *330*, 447–456.

(21) Borrelli, R.; Peluso, A. *J. Chem. Phys.* **2006**, *125*, 194308/1−8.

(22) Peluso, A.; Borrelli, R.; Capobianco, A. *J. Phys. Chem. A* **2009**, *113*, 14831–14837.

(23) Scalmani, G.; Frisch, M. J.; Menucci, B.; Tomasi, J.; Cammi, R.; Barone, V. *J. Chem. Phys.* **2006**, *124*, 094107/1−15.

(24) Furche, F.; Ahlrichs, R. *J. Chem. Phys.* **2004**, *121*, 12772/1−2.

(25) Köhn, A.; Hättig, C. *J. Chem. Phys.* **2003**, *119*, 5021/1−16.

(26) Kemper, M.; Van Dijk, J.; Buck, H. *Chem. Phys. Lett.* **1978**, *53*, 121–124.

(27) Berger, R.; Fischer, C.; Klessinger, M. *J. Phys. Chem. A* **1998**, *102*, 7157–7167.

(28) Dierksen, M.; Grimme, S. *J. Chem. Phys.* **2005**, *122*, 244101/1−9.

(29) Santoro, F.; Lami, A.; Improta, R.; Barone, V. *J. Chem. Phys.* **2007**, *126*, 184102/1−11.

(30) Jankowiak, H.-C.; Stuber, J. L.; Berger, R. *J. Chem. Phys.* **2007**, *127*, 234101/1−23.

(31) IDEA: In-Silico Developments for Emerging Applications; http://idea.sns.it. Accessed January 29, 2010.

(32) Eckart, C. *Phys. Rev.* **1937**, *47*, 552–558.

(33) Franck, J. *Trans. Faraday Soc.* **1926**, *21*, 536–542.

(34) Condon, E. U. *Phys. Rev.* **1928**, *32*, 858–872.

(35) Herzberg, G.; Teller, E. *Z. Phys. Chem. B-Chem. E* **1933**, *21*, 410–446.

(36) Luis, J. M.; Bishop, D. M.; Kirtman, B. *J. Chem. Phys.* **2004**, *120*, 813–822.

(37) Luis, J. M.; Torrent-Sucarrat, M.; Sola, M.; Bishop, D. M.; Kirtman, B. *J. Chem. Phys.* **2005**, *122*, 184104/1−13.

(38) Duschinsky, F. *Acta Physicochim. URSS* **1937**, *7*, 551–566.

(39) Sharp, T. E.; Rosenstock, H. M. *J. Chem. Phys.* **1963**, *41*, 3453–3463.

(40) Islampour, R.; Dehestani, M.; Lin, S. H. *J. Mol. Spectrosc.* **1999**, *194*, 179–184.

(41) Mebel, A. M.; Hayashi, M.; Liang, K. K.; Lin, S. H. *J. Phys. Chem. A* **1999**, *103*, 10674–10690.

(42) Kikuchi, H.; Kubo, M.; Watanabe, N.; Suzuki, H. *J. Chem. Phys.* **2003**, *119*, 729–735.

(43) Liang, J.; Li, H. *Mol. Phys.* **2005**, *103*, 3337–3342.

(44) Chang, J.-L. *J. Chem. Phys.* **2008**, *128*, 174111/1−10.

(45) Cederbaum, L. S.; Domcke, W. *J. Chem. Phys.* **1976**, *64*, 603–611.

(46) Doktorov, E. V.; Malkin, I. A.; Man'ko, V. I. *J. Mol. Spectrosc.* **1977**, *64*, 302–326.

(47) Ruhoff, P. T. *Chem. Phys.* **1994**, *186*, 355–374.

(48) Malmqvist, P.-Å.; Forsberg, N. *Chem. Phys.* **1998**, *228*, 227–240.

(49) Becke, D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(50) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51–57.

(51) Peach, M. J. G.; Helgaker, T.; Salek, P.; Keal, T. W.; Lutnaes, O. B.; Tozer, D. J.; Handy, N. C. *Phys. Chem. Chem. Phys.* **2006**, *8*, 558–562.

(52) Barone, V.; Cimino, P.; Stendardo, E. *J. Chem. Theory Comput.* **2008**, *4*, 751–764.

(53) Double and triple- $\zeta$ basis sets of N07 family, N07D, N07T and N07Tdiff; IDEA: In-Silico Developments for Emerging Applications; http://idea.sns.it. Accessed January 29, 2010.

(54) Barone, V. to be published.

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

(56) Kendall, R.; Dunning, T., Jr.; Harrison, R. *J. Chem. Phys.* **1992**, *96*, 6769–6806.

(57) Lin, N.; Santoro, F.; Zhao, X.; Rizzo, A.; Barone, V. *J. Phys. Chem. A* **2008**, *112*, 12401–12411.

(58) Barone, V.; Bloino, J.; Biczysko, M. *Phys. Chem. Chem. Phys.* **2010**, *12*, 1092–1101.

(59) Barone, V.; Biczysko, M.; Cimino, P. Interplay of stereo electronic vibrational and environmental effects in tuning physico-chemical properties of carbon centered radicals. In *Carbon-Centered Free Radicals and Radical Cations*; Forbes, M. D. E., Ed.; John Willey & Sons, Inc.: Hoboken, NJ, 2010; pp 105−139.

(60) Cossi, M.; Scalmani, G.; Rega, N.; Barone, V. *J. Comput. Chem.* **2003**, *24*, 669–681.

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

*Version, Revision H.05*, Gaussian, Inc.: Wallingford, CT, 2006.

(62) Grimme, S.; Diedrich, C. *J. Phys. Chem. A* **2003**, *107*, 2524–2539.

(63) Pecul, M.; Ruud, K.; Helgaker, T. *Chem. Phys. Lett.* **2004**, *388*, 110–119.

(64) Crawford, D. T.; Tam, M. C.; Abrams, M. L. *J. Phys. Chem. A* **2007**, *111*, 12057–12068.

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

(66) Barone, V.; Cimino, P. *Chem. Phys. Lett.* **2008**, *454*, 139–143.

(67) Barone, V.; Cimino, P. *J. Chem. Theory Comput.* **2009**, *5*, 192–199.

(68) Pecul, M.; Marchesan, D.; Ruud, K.; Coriani, S. *J. Chem. Phys.* **2005**, *122*, 024106/1−9.

(69) Vassiliev, S.; Bruce, D. *Photosynth. Res.* **2008**, *97*, 75–89.

(70) Du, H.; Fuh, R. A.; Li, J.; Corkan, A.; Lindsey, J. S. *Photochem. Photobiol.* **1998**, *68*, 141–142.

(71) Strain, H. H.; Thomas, M. R.; Katz, J. J. *Biochim. Biophys. Acta* **1963**, *75*, 306–311.

(72) Biczysko, M.; Borkowska, M.; Bloino, J.; Barone, V. in preparation.

(73) Tanaka, T.; Oelgemoller, M.; Fukui, F.; Aoki, F.; Mori, T.; Ohno, T.; Inoue, Y. *Chirality* **2007**, *19*, 415–427.

(74) Pescitelli, G.; Di Bari, L.; Caporusso, A. M.; Salvadori, P. *Chirality* **2008**, *20*, 393–399.

CT9006772

JCTC Journal of Chemical Theory and Computation

# Photoisomerization of Model Retinal Chromophores: Insight from Quantum Monte Carlo and Multiconfigurational Perturbation Theory

Omar Valsson* and Claudia Filippi*

*Faculty of Science and Technology and MESA+ Research Institute, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands*

Received December 26, 2009

**Abstract:** We present a systematic investigation of the structural relaxation in the excited state of model retinal chromophores in the gas phase using the complete-active-space self-consistent theory (CASSCF), multiconfigurational second-order perturbation theory (CASPT2), quantum Monte Carlo (QMC), and coupled cluster (CC) methods. In contrast to the CASSCF photoisomerization mechanism of bond inversion followed by torsion around formal double bonds, we find that the other approaches predict an initial skeletal relaxation which does not lead to bond inversion but to a rather flexible retinal chromophore with longer bonds and with the bond-length pattern of the ground state being partly preserved. The relaxation proceeds then preferentially via partial torsion around formal single bonds and does not reach a conical intersection region. Our findings are compatible with solution experiments which point to the existence of multiple minima and relaxation pathways, some of which are nonreactive, do not lead to photoproducts via conical intersection, and are dominant in solution. Our results also demonstrate the importance of a balanced description of dynamical and static correlation in the excited-state gradients and raise serious concerns on the common use of the CASSCF method to investigate structural properties of photoexcited retinal systems.

## 1. Introduction

The absorption of visible light and its conversion to other forms of energy is at the heart of some of the most fundamental processes in biology. An important example of light absorption initiating a biological response is the primary event of vision[1] where light induces the *cis*−*trans* isomerization of the photosensitive 11-*cis* retinal chromophore in rhodopsin[2] and other visual pigments, activating a cascade of chemical reactions which ultimately culminate in the stimulation of the optical nerve.[3] The initial photoisomerization process is one of the fastest photochemical reactions in nature, occurring within a few hundred femtoseconds,[4] and the protein environment plays a central role in guiding the reaction. In solution, the dynamics of retinal chromophores is in fact quite different than in the protein, namely, about 20 times slower[5] and much less efficient.[6,7] Even though femtosecond spectroscopy studies have extensively investigated the primary isomerization step of retinal chromophores,[8−14] the detailed nature of the molecular mechanism in the initial excited-state reaction and the exact role of the protein environment are still not understood.[12]

Theoretically, a large number of calculations with a variety of quantum chemical methods have been performed to investigate the structural and spectroscopic properties as well as the nature of the photoisomerization mechanism of retinal chromophores and retinal models in the gas phase[15−41] and in the protein environment.[24,42−61] Given the large size of the retinal chromophore, most calculations including the protein via quantum mechanics/molecular mechanics (QM/MM) approaches have mainly focused on obtaining a realistic representation of the structural model in the ground state and understanding the environmental effects on the absorption properties. Interestingly, even though all investigations employing different techniques appear to reproduce the correct experimental absorption value, the reasons behind

* Authors' e-mail: o.valsson@utwente.nl (O.V.); c.filippi@utwente.nl (C.F.).

10.1021/ct900692y © 2010 American Chemical Society
Published on Web 03/12/2010

this agreement are contrasting and there are fundamental differences concerning the structure of the chromophore, the protonation of nearby residues, and the overall role of the protein in tuning the spectral properties.[46,52,53,57,60] If nailing down the exact role of the environment on the Franck−Condon excitation has proven elusive, the computation of dynamical properties of photoexcited retinal chromophores in the gas phase as well as in the protein is an even harder task since it requires a uniformly reliable computation of excited-state potential energy surfaces and the availability of analytical energy gradients for geometric optimization and dynamical runs.

To date, most excited-state geometrical investigations have employed the low-level complete-active-space self-consistent field (CASSCF) approach for the relaxation of retinal chromophores in the gas phase[15−18,33,34,38,46] with also few attempts to simulate the dynamics of retinal chromophores in the protein environment.[45,55,61] The excited-state energies on the CASSCF structures are often refined in single-point calculations with higher-level approaches such as multiconfigurational perturbation theory (CASPT2) to partially account for dynamical correlation largely missing in the CASSCF description. These studies have led to the generally accepted picture that photoisomerization begins with an in-plane skeletal relaxation which yields bond inversion and proceeds via a torsional motion around carbon−carbon bonds having double-bond character in the ground state.[16,18,34,38] The chromophore is then funneled into a conical intersection region which leads to the ground-state *trans* photoproduct. Recent calculations of Send and Sundholm based on coupled cluster (CC) theory have however challenged this picture, as they obtained a rather different excited-state relaxation mechanism of retinal models in the gas phase.[29,31,35,37] The initial relaxation in the excited state at the CC level does not yield bond inversion but the lengthening of most bonds while preserving the general bond-length pattern of the ground state. The subsequent torsional motion is around carbon−carbon bonds holding single-bond character in the ground state. However, these CC calculations have been dismissed on the basis of being single reference and thus lacking a proper description of static correlation as compared to the CASSCF approach.[36] This response reflects the general acceptance of CASSCF as an adequate tool for the investigation of excited-state structural properties of retinal and other photosensitive chromophores.

In the present work, we perform a thorough investigation of the initial excited-state relaxation from the Franck−Condon point of model retinal chromophores in the gas phase and employ CASPT2 and quantum Monte Carlo (QMC) in addition to the CASSCF and CC methods. The CASPT2 approach is well established and is considered a benchmark method for the computation of excited-state properties, but its use for retinal models has been mostly confined to single-point calculations, in-plane geometrical relaxation of few models,[20] and constraint optimization of a minimal chromophore model.[19,20,23,41] The QMC method is instead less common in the field of theoretical photochemistry, and its use for excited-state geometrical optimization is novel. QMC has recently given promising results in the study of various photoactive molecules,[62−66] and a favorable comparison with CASPT2 will further establish its use for investigating photochemical problems. At the cost of being computationally more expensive, CASPT2 and QMC methods can give an accurate and balanced description of both static and dynamical correlation and therefore represent an ideal tool to clarify the nature of the microscopic mechanism in the photoisomerization of retinal chromophores and to resolve the disagreement between the generally accepted CASSCF picture and the recent, controversial CC results.

We find that our in-plane geometrical relaxations from the Franck−Condon point of retinal models show a consistent, good agreement among the CASPT2, CC, and QMC approaches, which give excited-state structures not characterized by bond-length inversion, and are in striking contrast to the results obtained with the CASSCF approach. Photoexcitation therefore weakens all bonds, which stretch and become partly more similar in length while preserving the general bond-length pattern of the ground state. To investigate a nontrivial minimum energy path out of plane, we consider a model with four double bonds and find that the excited-state relaxation at the CASPT2 level proceeds preferentially via a partial torsional motion around a formal single bond and does not lead to a conical intersection region. To assess the existence of a reactive path at the CASPT2 level, we also study the constrained excited-state isomerization around a formal double bond for the same model. We find that the system encounters a small barrier to isomerization at rather large angles of rotation, beyond which it is funneled toward a conical intersection region characterized by bond inversion.

Therefore, in agreement with previous CC calculations, our results support the picture of a rather flexible retinal chromophore in the excited state as compared to the CASSCF excited chromophore, which can only twist around formal double bonds. These findings are compatible with the observation in solution experiments of the existence of multiple minima, possibly corresponding to different torsional conformations, and multiple excited-state paths. Some of these paths are reactive and yield a photoproduct via a conical intersection, while others are nonreactive, do not lead to a conical intersection, and are dominant in solution.[67] Finally, our results demonstrate the importance of including an accurate description of dynamical correlation also in the excited-state gradients and raise serious concerns about the common use of CASSCF in investigating excited-state structures of retinal systems.

In section 2, we briefly present the methods used in this paper and focus on the description of the QMC geometrical optimization. In section 3, we describe the computational details, and in section 4, we introduce the model retinal chromophores we investigate. In sections 5−7, we present the results for the vertical excitation energy, the in-plane geometrical relaxation, and the minimum energy path or out-of-plane geometrical relaxation in the excited state. Finally, in section 8, we discuss our results and conclude.

## 2. Methods

In this work, we employ a wide range of ab initio quantum chemical methods. While we refer the reader to appropriate textbooks[68] for a discussion of the more traditional CASSCF, CASPT2, and CC approaches, we briefly review below the less common QMC methods.[69] In particular, we focus on the procedure we follow to perform geometrical optimization within variational Monte Carlo (VMC), which is nonstandard, and on how we address stability issues in the calculation of energy gradients.

**2.1. QMC Methods.** QMC methods provide an accurate and balanced description of dynamical and static electronic correlation in both molecular and extended systems.[69] Their application to the description of the excited-state properties of photoactive molecules has already given promising results.[62−66]

A crucial ingredient which determines the quality of a QMC calculation is the many-body trial wave function, which is here chosen to be of the so-called Jastrow-Slater type. Since we treat multiple states of the same symmetry, we write the ground- and excited-state wave functions as a linear combination of spin-adapted configuration state functions (CSF) multiplied by a Jastrow correlation factor:

$$\psi_I = \mathcal{J}\sum_{i=1}^{N_{\mathrm{CSF}}} c_i^I C_i \tag{1}$$

where different states depend on their individual linear coefficients, $c_i^I$, but share a common set of single-particle orbitals and Jastrow factor, $\mathcal{J}$. We use here a Jastrow factor which correlates pairs of electrons and each electron separately with a nucleus, and we employ different Jastrow factors to describe the correlation with different atom types. Since the optimal orbitals and expansion coefficients in $\psi_I$ may differ from the values obtained for instance in a CASSCF calculation in the absence of the Jastrow factor, it is important to reoptimize them in the presence of the Jastrow factor.

The parameters of the trial wave functions are optimized in an efficient and simple approach in a state-average (SA) fashion as described in ref 66. In this scheme, we iteratively alternate between optimizing the linear coefficients in the CSF expansion and the nonlinear (Jastrow and orbital) coefficients where the quantity minimized is the weighted averaged energy over the states under consideration:

$$E_{\mathrm{SA}} = \sum_I w_I \frac{\langle \Psi_I|\mathcal{H}|\Psi_I\rangle}{\langle \Psi_I|\Psi_I\rangle} \tag{2}$$

where the weights are fixed and $\sum_I w_I = 1$. At convergence, the averaged energy $E_{\mathrm{SA}}$ is stationary with respect to all parameter variations subject to the orthogonality constraint, while the energies of the states are stationary with respect to variations of the linear coefficients but not of the orbital or Jastrow parameters. In this approach, the wave functions are kept orthogonal and a generalized variational theorem applies.

The set of optimal linear coefficients is obtained by solving a generalized eigenvalue problem where the Hamiltonian and the overlap matrix on the basis functions $\mathcal{J}C_i$ are estimated within VMC by sampling a guiding function $\Psi_g$ chosen to have significant overlap with all states of interest. The use of a nonsymmetric estimator of the Hamiltonian matrix yields a strong zero-variance principle and renders the approach particularly efficient.[70] To optimize the nonlinear parameters, we employ the linear optimization method first developed for ground states[71] and recently extended to the state-average optimization of multiple states.[66] In this scheme, the nonlinear minimization problem is linearized by working on the basis of the derivatives of the wave function with respect to the nonlinear parameters. In the case of multiple states, the elements of the weighted averaged Hamiltonian and overlap matrices are computed in a single VMC run by sampling a guiding wave function $\Psi_g$. When determining both the linear and the nonlinear parameters, the guiding wave function is here chosen equal to $\sqrt{(\Sigma_I|\Psi_I|^2)}$.

The optimal trial wave functions are then used in diffusion Monte Carlo (DMC), which gives the best energy within the fixed-node approximation, that is, the lowest-energy state with the same zeros (nodes) as the trial wave function.

**2.2. VMC Geometrical Optimization.** The VMC geometrical optimization is performed in Z-matrix coordinates where the energy gradients with respect to the nuclear coordinates are obtained using numerical differentiation and correlated sampling.[72]

To determine the interatomic forces at a given reference geometry, we construct a set of secondary geometries corresponding to small forward and backward displacements of 0.001 au for the bond lengths and 0.01° for the bond and dihedral angles. The gradient in Z-matrix coordinates is computed as

$$\mathbf{g}_\gamma = [E(\mathbf{x} + \delta x_\gamma) - E(\mathbf{x} - \delta x_\gamma)]/2\delta x_\gamma \tag{3}$$

where $E$ is the total energy and $\delta x_\gamma$ is a displacement in the internal coordinate $\gamma$ with respect to the reference coordinates $\mathbf{x}$. The diagonal component of the Hessian can be obtained in the same run at no extra cost as

$$\mathbf{h}_\gamma^{\mathrm{diag}} = [E(\mathbf{x} + \delta x_\gamma) - 2E(\mathbf{x}) + E(\mathbf{x} - \delta x_\gamma)]/\delta x_\gamma^2 \tag{4}$$

The geometry is updated according to an approximate version of the Newton−Raphson method as

$$\mathbf{x}'_\gamma = \mathbf{x}_\gamma - \mathbf{g}_\gamma/\mathbf{h}_\gamma^{\mathrm{diag}} \tag{5}$$

where $\mathbf{x}'$ denotes the new coordinates in Z-matrix representation. To stabilize the procedure against numerical noise, we add a constant parameter of $5 \times 10^{-5}$ to all diagonal elements of the Hessian.

The use of correlated sampling allows us to efficiently determine relative energies for different geometries from a single reference Monte Carlo walk. The reference walk is obtained by sampling the wave function $\Psi$ corresponding to the coordinates $\mathbf{x}$ and Hamiltonian $\mathcal{H}$, while the secondary geometries are characterized by the corresponding quantities $\mathbf{x} \pm \delta x_\gamma$, $\Psi_\gamma$, and $\mathcal{H}_\gamma$. Given a reference primary wave function, the secondary wave function is here simply obtained by recentering $\Psi$ at the coordinates $\mathbf{x} \pm \delta x_\gamma$ without altering the wave function parameters. The electronic coordinates of the secondary walk are obtained by stretching the primary walk with the nuclear coordinates through a space-warp transformation as described in ref 72. In the present work, we use the function $F(r) = r^{-4}$ for the space-warp transformation.

In summary, the procedure for the geometrical optimization is the following: (*i*) The determinantal component of the initial wave function is obtained in a CASSCF calculation. (*ii*) All wave function parameters are optimized in a VMC run (we discuss later the importance of optimizing the orbital parameters). (*iii*) The energy gradients are obtained in a correlated sampling VMC calculation. (*iv*) The geometry is updated as described above. We note that, with the exception of the first iteration, step *i* can be skipped since step *ii* can be performed starting from the wave function optimized at the previous geometry and recentered at the current geometry. This procedure is iterated until the bond length and bond angle gradients are on the order of 0.001 hartree/Bohr and 0.0001 hartree/deg, respectively, that is, comparable to their error bars. Since the stochastic nature of VMC does not allow the assignment of one particular geometry as the minimum one, we perform 5−10 additional iterations after convergence and average the internal coordinates over these additional steps.

To decrease the computational effort, the carbon−hydrogen and nitrogen−hydrogen bond lengths and all the bond angles involving terminal hydrogen atoms are kept fixed. All other internal degrees of freedom are allowed to vary.

**2.3. Stability of VMC Energy Gradients.** The computation of gradients in VMC poses some stability issues which we analyze by considering for simplicity the gradient expression without the use of the space-warp transformation. Then, the energy difference between the primary and a secondary geometry can be written as

$$E(\mathbf{x}) - E(\mathbf{x} + \delta x_\gamma) = \left\langle \frac{\mathcal{H}\Psi(\mathbf{R})}{\Psi(\mathbf{R})} - \frac{\mathcal{H}_\gamma \Psi_\gamma(\mathbf{R})}{\Psi_\gamma(\mathbf{R})} W_\gamma(\mathbf{R}) \right\rangle_{\Psi^2} \qquad (6)$$

where $\langle \cdots \rangle$ denotes the statistical average over the configurations sampled in VMC from the distribution $\Psi^2$, and the weights are defined as

$$W_\gamma(\mathbf{R}) = \frac{\Psi_\gamma^2(\mathbf{R})/\Psi^2(\mathbf{R})}{\langle \Psi_\gamma^2(\mathbf{R})/\Psi^2(\mathbf{R}) \rangle_{\Psi^2}} \qquad (7)$$

If we expand this energy difference to linear order in $\delta x_\gamma$, we obtain a term proportional to

$$\left\langle \frac{\mathcal{H}\Psi(\mathbf{R})}{\Psi(\mathbf{R})} \frac{\partial \log \Psi(\mathbf{R})}{\partial x_\gamma} \right\rangle_{\Psi^2} \qquad (8)$$

Since the product inside the square brackets diverges as $1/d^2$ when the distance $d$ from the nodes of $\Psi$ approaches zero, the estimator of eq 6 obtained by sampling the square of the primary wave function has infinite variance, and it is not possible to obtain a stable energy difference. To cure this problem, we follow ref 73 and sample a different distribution which is nonzero at the nodes and is defined here as

$$\Psi_{\mathrm{g}}(\mathbf{R}) = \Psi(\mathbf{R}) \frac{\max[\varepsilon, d_{\mathrm{n}}(\mathbf{R})]}{d_{\mathrm{n}}(\mathbf{R})} \qquad (9)$$

where $d_{\mathrm{n}}(\mathbf{R})$ is a measure of the distance from the nodes:

$$d_{\mathrm{n}}(\mathbf{R}) = \frac{1}{|\nabla\Psi(\mathbf{R})/\Psi(\mathbf{R})|} \qquad (10)$$

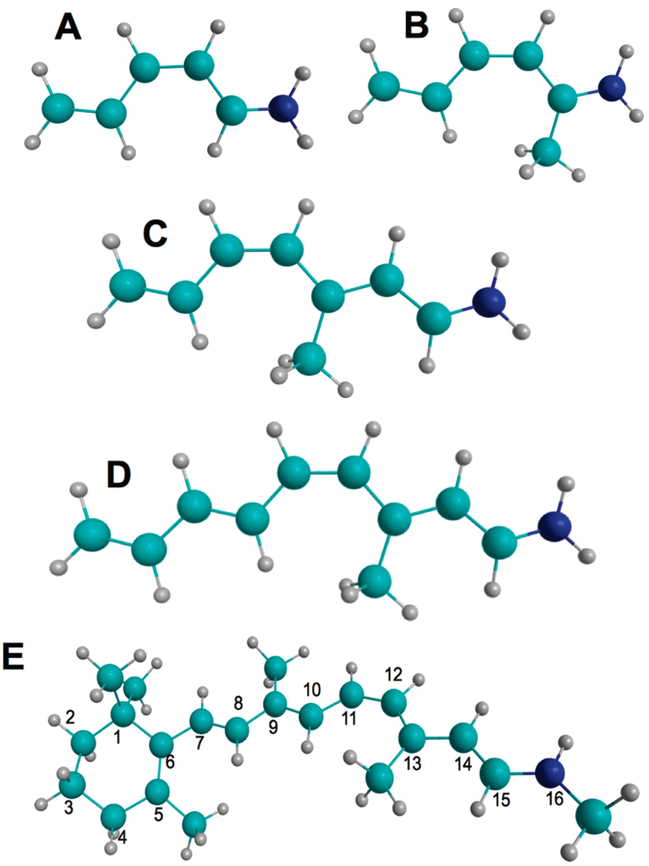


***Figure 1.*** Model retinal chromophores. The atom numbering for chromophore E is used for all models, so the *cis* bond is always between $C_{11}$ and $C_{12}$. Cyan, blue, and gray represent carbon, nitrogen, and hydrogen, respectively.

and $\varepsilon$ is a cutoff parameter[74] chosen as $10^{-2}$. The average of eq 8 can then be rewritten as

$$\left\langle \frac{\Psi^2(\mathbf{R})}{\Psi_{\mathrm{g}}^2(\mathbf{R})} \frac{\mathcal{H}\Psi(\mathbf{R})}{\Psi(\mathbf{R})} \frac{\partial \log \Psi(\mathbf{R})}{\partial x_\gamma} \right\rangle_{\Psi_{\mathrm{g}}^2} \qquad (11)$$

where the reweighting factor $\Psi^2(\mathbf{R})/\Psi_{\mathrm{g}}^2(\mathbf{R})$ removes the divergence of the products inside the brackets. This cures the problem of the infinite variance and allows us to obtain stable energy differences.

## 3. Computational Details

We use the program MOLCAS 7.2[75] to optimize the ground-state geometries of the model chromophores within all-electron MP2 and DFT with the B3LYP[76] functional. For the ground-state optimizations of the full retinal model (see Figure 1E), we employ the Gaussian 03 code.[77] The default convergence criteria are used for both codes.

We also use MOLCAS 7.2 for the all-electron CASSCF, CASPT2, and multistate (MS) CASPT2[78] calculations. The state-average (SA) CASSCF calculations are performed with equal weights over the states of interest, and the two lowest states are used in the SA-CASSCF and MS-CASPT2 calculations. In the CASPT2 calculations, we employ the default IPEA zero-order Hamiltonian[79] unless otherwise stated and indicate if an additional constant level shift[80] is added to the Hamiltonian. In the CASPT2 calculations for the complete 11-*cis* retinal chromophore, we use the Cholesky decomposition of the two-electron integrals[81] with a default threshold of $10^{-4}$. Analytical CASSCF and numerical CASPT2 gradients are used for geometrical optimizations and minimum energy path (MEP) calculations. In the CASPT2 calculations, we do not correlate as many lowest orbitals of $\sigma$ character as the number of heavy atoms in the model. The default convergence criteria are used for all calculations.

The EOM-CC calculations are performed with the codes ACES II[82] and CFOUR.[83] The CC calculations include approximate single and double excitations (CC2) and single and double excitations (CCSD). Default convergence criteria are used for all calculations, and we do not correlate as many lowest orbitals of $\sigma$ character as the number of heavy atoms in the model.

The program package CHAMP[84] is used for the QMC calculations. We employ scalar-relativistic energy-consistent Hartree−Fock pseudopotentials[85] where the carbon, nitrogen, and oxygen 1s electrons are replaced by a nonsingular s-nonlocal pseudopotential and the hydrogen potential is softened by removing the Coulomb divergence. Different Jastrow factors are used to describe the correlation with different atom types, and for each atom type, the Jastrow factor consists of an exponential of the sum of two fifth-order polynomials of the electron−nuclear and the electron−electron distances, respectively.[86] We also test the effect of including an electron−electron−nuclear term. The starting determinantal components are obtained in CASSCF calculations, which are performed with the program GAMES-S(US).[87] In all SA-CASSCF calculations, equal weights over the states are employed and the final CAS expansions are expressed on the weighted-average CASSCF natural orbitals. The CAS wave functions of the ground and excited states may be truncated with an appropriate threshold on the CSF coefficients, and the union sets of surviving CSFs for the states of interest are retained in the QMC calculations. The Jastrow correlation factor and the CI coefficients are optimized by energy minimization in a state-averaged sense within VMC with equal weights. When indicated in the text, also the orbitals are optimized along with the Jastrow and CI parameters. The pseudopotentials are treated beyond the locality approximation,[88] and an imaginary time step of 0.05 or 0.075 au is used in the DMC calculations.

In the DFT, CASSCF, CASPT2, and CC calculations, we employ either the correlation consistent (cc-pVxZ)[89] or the atomic natural orbital (ANO-L-VxZP)[90] basis sets. We use the ANO contraction schemes as defined in MOLCAS, that is, [3s2p1d]/[2s1p] for ANO-L-VDZP, [4s3p2d1f]/[3s2p1d] for ANO-L-VTZP, and [5s4p3d2f]/[4s3p2d] for ANO-L-VQZP. In the single-point energy calculations, the ANO-L-VDZP basis set is generally used, while the default basis in the geometrical optimization and MEP calculations is the cc-pVDZ. In the QMC calculations, we use the Gaussian basis sets[85] specifically constructed for our pseudopotentials. In particular, we employ the cc-pVDZ basis (denoted by D) and the D basis augmented with s and p diffuse functions[91] on the heavy atoms (denoted by D+). We also use the T′ and Q′ basis sets which consist of the cc-pVTZ and cc-pVQZ, respectively, combined with the cc-pVDZ for hydrogen. The g functions are not included in the Q′ basis. Most single-point energy calculations use the D+ basis, while geometrical optimizations employ the D basis.

## 4. Retinal Models

The 11-*cis* retinal chromophore consists of a conjugated carbon chain with a protonated Schiff base (PSB) at one end and a twisted $\beta$-ionone ring at the other end (see Figure 1E). It sits inside the protein pocket of rhodopsin, a seven helix transmembrane protein, where it is covalently bound to Lys-296 via the protonated Schiff base linkage. In theoretical gas phase studies, there has been no consistent choice of how to terminate the covalent bond between the positively charge nitrogen in the protonated Schiff base and Lys-296. A single hydrogen, a methyl, and also an *n*-butyl group have been used as termination, and this particular choice appears to influence only slightly the excitation energy.[27,40] Due to the large size of the 11-*cis* retinal chromophore, smaller protonated Shiff base models have been mainly investigated theoretically, which differ in the length of the conjugated chain and the absence of methyl groups with respect to the complete chromophore.

The retinal models studied in this work are shown in Figure 1 and range from the minimal model (Figure 1A) to the full 11-*cis* retinal chromophore (Figure 1E). The atom labeling shown for the 11-*cis* chromophore is adopted also for the other models so that the *cis*-to-*trans* isomerization bond is always between the $C_{11}$ and $C_{12}$ atoms, with atom numbering increasing from the carbon to the nitrogen end. For the models without the $\beta$-ionone ring, we introduce the naming convention PSBx(y) where x and y are the number of double bonds and methyl groups, respectively. The PSB3(0) ($C_5H_6NH_2^+$) model (A) is the minimal model of the retinal chromophore and has already been extensively studied in the literature.[15,17,20,22,25,41] Since the methyl group at position $C_{13}$ plays an important role in accelerating the isomerization process,[15,17] we also consider the PSB3(1) ($C_6H_8NH_2^+$) model (B), that is, the minimal model (A) with an added methyl group. The PSB4(1) ($C_8H_{10}NH_2^+$) model (C) has one additional double bond and has been previously studied without the methyl group using the CC and TDDFT methods.[29] The PSB5(1) ($C_{10}H_{12}NH_2^+$) model (D) has the full conjugated chain but is missing the $\beta$-ionone ring, and the complete 11-*cis* retinal chromophore (E) is here terminated with a single methyl group. With the exception of the 11-*cis* (E) chromophore, all other models are planar in the ground state. We note that a direct comparison with experiments is only possible for the vertical excitation energy of the 11-*cis* chromophore (E) since, to the best of our knowledge, none of the smaller models has been studied experimentally.

## 5. Vertical Excitation Energies

We compute the vertical excitation energies of the lowest singlet excited state ($S_1$) of all retinal models using the CASPT2, CC2, CCSD, VMC, and DMC approaches. The ground-state DFT/B3LYP geometries optimized with the cc-

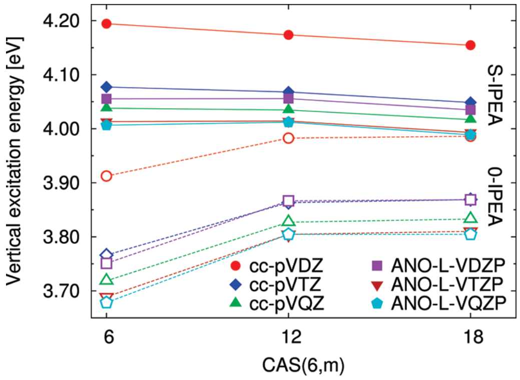


***Figure 2.*** MS-CASPT2 vertical excitation ($S_1$) energies of the PSB3(0) model (A) computed with the standard IPEA Hamiltonian (S-IPEA, filled symbols) and with the IPEA shift set to zero (0-IPEA, empty symbols). The excitations are obtained with different basis sets and expansions CAS(6, *m*) of 6 electrons in *m* active orbitals. The ground-state DFT/B3LYP geometry is used.

***Table 1.*** VMC and DMC Vertical Excitation ($S_1$) Energies (eV) of the PSB3(0) Model (A), Computed with Different Basis Sets and CAS Expansions Expressed on the Weighted-Averaged Natural Orbitals[a]

| CAS(6, *m*) | Thr. | Det./CSF | Jastrow | basis | VMC | DMC |
|---|---|---|---|---|---|---|
| (6,6) | 0.01 | 183/79 | e−n, e−e | D+ | 4.32(1) | 4.22(2) |
| (6,6) | 0.02 | 101/47 | e−n, e−e | D+ | 4.31(1) | 4.20(2) |
| (6,6) | 0.04 | 66/31 | e−n, e−e | D+ | 4.31(1) | 4.21(2) |
| (6,6) | 0.08 | 23/10 | e−n, e−e | D+ | 4.24(2) | 4.19(2) |
| (6,6)[b] | 0.08 | 23/10 | e−n, e−e | D+ | 4.25(2) | 4.21(2) |
| (6,6)[c] | 0.08 | 23/10 | e−n, e−e | D+ | 4.28(1) | 4.16(2) |
| (6,6) | 0.02 | 103/48 | e−n, e−e | D | 4.38(1) | 4.29(2) |
| (6,6) | 0.02 | 103/48 | e−n, e−e | T′ | 4.34(1) | 4.25(2) |
| (6,6) | 0.02 | 103/48 | e−n, e−e | Q′ | 4.34(1) | 4.22(2) |
| (6,12) | 0.02 | 152/66 | e−n, e−e | D+ | 4.29(1) | 4.22(2) |
| (6,18) | 0.02 | 156/67 | e−n, e−e | D+ | 4.29(1) | 4.22(2) |
| (6,6) | 0.02 | 101/47 | e−n, e−e, e−e−n | D+ | 4.32(2) | 4.24(2) |

[a] The CAS(6,*m*) active space includes all 6 $\pi$ electrons occupied in the reference configuration and *m* active $\pi$ orbitals. The threshold on the expansion and the corresponding number of determinants and CSFs are also listed. Unless indicated, only the Jastrow and CI parameters are optimized. The ground-state DFT/B3LYP geometry is used. [b] Orbitals optimized including 40 external orbitals. [c] Orbitals optimized including 80 external orbitals.

pVDZ basis are used. The model E is optimized with no symmetry constraint ($C_1$), while the other models are planar and are optimized in either $C_s$ or $C_1$ symmetry. The CASPT2 excitations are computed with the standard IPEA Hamiltonian (S-IPEA) and with the IPEA shift set to zero (0-IPEA), which was the default prior to MOLCAS 6.4, in order to be compatible with previous calculations in the literature.

**5.1. Dependence on Basis Set and Other Parameters.** Before comparing the relative performance of the different methods, we focus on the minimal model (A) and investigate the dependence of the excitations on the choice of the basis set and other parameters which may affect the calculations. We begin with the MS-CASPT2 approach and show in Figure 2 the vertical excitations obtained with the double (D), triple (T), and quadruple (Q) $\zeta$ basis sets from the cc-pVxZ and ANO-L-VxZP series. We correlate all 6 $\pi$ electrons in the reference configuration and use a different number *m* of virtual $\pi$ orbitals in the CAS(6,*m*) expansion. We note that single-state and MS-CASPT2 yield equivalent excitations for model A.

We observe that the ANO-L-VxZP series gives a faster convergence in the CASPT2 excitation energy than the correlated consistent basis. The excitations computed with the D basis are only 0.05 eV higher than the values obtained with the T and Q basis sets. On the other hand, in the cc-pVxZ series, the D excitations are 0.12 eV higher than the T values, which still differ from the Q results by 0.04 eV. The behavior of the CC2 and CCSD excitations with the basis set is not shown in the figure but parallels closely the one observed for the CASPT2 excitations. Therefore, since the ANO-L-VDZP basis set gives a good balance between accuracy and computational cost, it is used hereafter for all single-point CASPT2, CC2, and CCSD calculations.

We find that the CASPT2 results depend very strongly on the choice of the zero-order Hamiltonian. The difference between the excitation energies obtained with the standard IPEA Hamiltonian and the IPEA shift set to zero is independent of the basis set used and equal to about 0.3 eV when a CAS(6,6) is employed. As expected, the dependence on the particular zero-order Hamiltonian is reduced as the wave function is improved, and the difference between the excitations with and without the IPEA shift becomes 0.2 eV if the number of active $\pi$ orbitals in the CAS is increased from 6 to 18. Finally, we observe that the vertical energies obtained with the IPEA shift set to zero are much more sensitive to the dimension of the CAS since they increase by 0.07−0.12 eV when *m* goes from 6 to 18, while the energies obtained with the standard IPEA Hamiltonian are quite stable and only decrease by about 0.02−0.04 eV.

In Table 1, we present an extensive QMC investigation for the minimal model (A) to assess how different ingredients in the trial wave function affect the excitation energy. The reference wave function is constructed from a CAS(6,6) expansion expressed on the weighted-averaged CASSCF natural orbitals in the D+ basis and truncated with a threshold of 0.02, where only the two-body Jastrow factor and CI coefficients are optimized in energy minimization in a SA fashion. Starting from this wave function, we investigate the effect of (*i*) changing the threshold on the CAS(6,6) expansion in the range 0.01−0.08, (*ii*) increasing the number of active $\pi$ orbitals from 6 to 18 in the CAS(6,*m*) expansion, (*iii*) including an electron−electron−nuclear (e−e−n) term in the Jastrow factor in addition to the electron−nucleus (e−n) and electron−electron (e−e) components, (*iv*) optimizing the orbitals in a CAS(6,6) wave function with a threshold of 0.08 with both 40 and 80 external orbitals included in the optimization, and (*v*) using different basis sets (D, T′, and Q′). We find that the choice of basis has a significant effect on the QMC results as the VMC and DMC excitations computed with the D basis are higher by 0.06(2) and 0.09(3) eV than the corresponding D+ values. Since the use of the larger T′ and Q′ basis sets brings the excitations in closer agreement with the D+ results, we employ below the D+ basis set to compute the QMC excitations of all model chromophores. For this choice of basis, other ingredients in the trial wave function appear to have a smaller effect on the VMC and DMC excitation energies which range between 4.24(2) −4.32(2) and 4.16(2)−4.24(2) eV, respectively.

***Table 2.*** Vertical Excitation ($S_1$) Energies (eV) of the Retinal Models[a]

| | | MS-CASPT2 | | | | | |
|---|---|---|---|---|---|---|---|
| model | *n* | 0-IPEA | S-IPEA | CC2 | CCSD | VMC | DMC |
| (A) PSB3(0) | 6 | 3.75 | 4.06 | 4.12 | 4.23 | 4.31(1) | 4.20(2) |
| (B) PSB3(1) | 6 | 3.86 | 4.18 | 4.20 | 4.37 | 4.52(2) | 4.42(2) |
| (C) PSB4(1) | 8 | 3.04 | 3.35 | 3.33 | 3.47 | 3.59(2) | 3.47(2) |
| (D) PSB5(1) | 10 | 2.58 | 2.87 | 2.82 | 2.95 | 3.08(2) | 3.00(3) |
| (E) 11-*cis* | 12 | 2.03[b] | 2.30 | | | 2.59(3) | 2.41(3) |

[a] The MS-CASPT2 energies are computed both with the standard IPEA Hamiltonian (S-IPEA) and without the IPEA shift (0-IPEA). The CAS(*n*,*n*) expansion in the CASPT2 and QMC calculations includes all $\pi$ electrons in the reference configuration and an equal number *n* of $\pi$ orbitals. CASPT2 and CC employ the ANO-L-VDZP basis, and QMC the D+ basis. The ground-state DFT/B3LYP geometries are used. [b] Constant level shift of 0.2 au.

**5.2. Results.** We collect the vertical excitations of all retinal models computed using the MS-CASPT2, CC2, CCSD, VMC, and DMC methods on the ground-state DFT/B3LYP geometries in Table 2. The VMC and DMC excitations are obtained using wave functions where the Jastrow and CI parameters are optimized by energy minimization in a SA fashion and the threshold on the CSF expansion is 0.08 for the E model and 0.02 for all other models. It is evident that, for all models, the CASPT2 excitations obtained with the IPEA shift set to zero are at variance and significantly lower than the results obtained with all other theoretical methods. The use of the standard IPEA Hamiltonian raises the excitation energies of all models by as much as 0.3 eV and brings the CASPT2 values in close agreement with the CC2 results. The CCSD method yields excitations slightly higher by 0.11−0.17 eV than the CC2 and CASPT2 results obtained with the IPEA Hamiltonian. Finally, the VMC excitations are always higher by 0.1−0.2 eV than the DMC values which agree closely with the CCSD results.

For a comparison with experiments and previous theoretical work, we focus on the full 11-*cis* retinal chromophore (E) and collect the relevant results in Table 3. In line with previous calculations,[24,30,58] we find that the excitation energy of the retinal chromophore depends strongly on the method used to determine its ground-state structure. The sequence of BLYP, B3LYP, MP2, and CASSCF geometries corresponds to an increase of the degree of bond-length alternation and of the twisting of the $\beta$-ionone ring from −30° to −60° (see Figure 3). Stronger bond alternation and larger twisting angles correspond to larger excitations energies, and we find indeed an increase of 0.5 eV in the CASPT2 excitation both with and without the IPEA shift, when going from the BLYP to the CASSCF geometry. A comparison with CASPT2 geometries of planar retinal models indicates that DFT and MP2 ground-state structures represent a better model for the retinal chromophore in the gas phase as shown in Figure 6a and Figure SI-5 (Supporting Information) and already observed in ref 20. Even though discarding the CASSCF structures significantly reduces the spread in excitations, we still have an uncertainty of about 0.1 eV related to the choice of the particular DFT or MP2 geometry.

In Table 3, we present the single-state (SS) excitations in addition to results obtained with the MS-CASPT2 approach as done so far in this section. As already mentioned, SS-CASPT2 and MS-CASPT2 give equivalent excitations within 0.01 eV for the smaller model A, as expected given the large gap of about 4 eV between the ground and excited states. However, as the size of the retinal model increases and the excitation decreases, SS-CASPT2 and MS-CASPT2 start to differ, and this discrepancy grows faster when no IPEA shift is employed. For the 11-*cis* model (E) and a gap of about 2 eV, the difference amounts to about 0.10 and 0.15 eV with and without the IPEA shift, respectively, and is independent of the ground-state geometry. Therefore, the choice of performing single- or multistate calculations within CASPT2 represents another internal parameter of the theory which affects the CASPT2 excitation in addition to the IPEA shift. We remark that, while MS-CASPT2 gives results which nicely parallel the DMC and CC excitations for all models, the difference between CASPT2 and other theories increases with system size if the single-state approach is used. The choice of the MS theory is our preference also for compat-

***Table 3.*** Single-State (SS) and MS-CASPT2, and DMC Vertical Excitations ($S_1$) Energies (eV) of the 11-*cis* Retinal (E) Chromophore[a]

| Method | Geometry | $E_{exc}$ | |
|---|---|---|---|
| SS-CASPT2 | | 0-IPEA | S-IPEA |
| | DFT/BLYP | 1.81[c] | 2.12 |
| | DFT/B3LYP | 1.89[c] | 2.20 |
| | MP2 | 1.92[c] | 2.24 |
| | CASSCF[b] | 2.30[c] | 2.65[c] |
| MS-CASPT2 | | | |
| | DFT/BLYP | 1.96[c] | 2.22 |
| | DFT/B3LYP | 2.03[c] | 2.30 |
| | MP2 | 2.08[c] | 2.35 |
| | CASSCF[b] | 2.42[c] | 2.72[c] |
| DMC/D+ | | | |
| | DFT/BLYP | 2.32(3) | |
| | DFT/B3LYP | 2.41(3) | |
| Expt.[92] | | 2.05−2.34[d] | |

[a] The experimental estimate is also listed. The geometries are optimized with the cc-pVDZ basis, and the CASPT2 calculations employ a CAS(12,12) expansion and the ANO-L-VDZP basis. [b] CASSCF(12,12)/6-31G(d) geometry from ref 27. [c] Constant level shift of 0.2 au. [d] Termination with two methyl groups, $-N(CH_3)_2^+$.

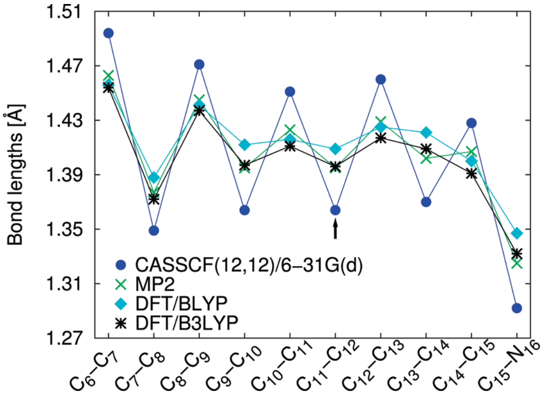


***Figure 3.*** Ground-state bond lengths (Å) of the 11-*cis* chromophore (E) optimized using MP2, DFT/BLYP, and B3LYP and the cc-pVDZ basis. The CASSCF(12,12)/6-31G(d) geometry is from ref 27. The $C_5-C_6-C_7-C_8$ dihedral angles are −29.7°, −33.5°, −40.5°, and −68.8° in BLYP, B3LYP, MP2, and CASSCF, respectively.

ibility with the CASPT2 excited-state geometrical optimizations presented in the next sections, where we employ the MS approach, as it is not known a priori whether the molecule will encounter a conical intersection region during relaxation.

We now compare our theoretical results with gas phase photodestruction experiments which are available for the 11-*cis* model terminated with two methyl groups.[92] The experimental absorption spectrum displays two main peaks at 2.05 eV (610 nm) and 3.18 eV (390 nm), which have been interpreted as the location of the vertical excitations to the two lowest singlet excited states ($S_1$ and $S_2$). The lowest-energy band ($S_1$) displays however a broad shoulder which has a secondary peak at 2.34 eV (530 nm) and is only about 20% lower in intensity than the absorption maximum at 2.05 eV. It has been previously suggested[33] that the vertical transition lies in the broad shoulder at higher energies and corresponds to the secondary peak at 2.34 eV. We further note that the adiabatic and not the vertical transition may be related to the lowest-energy feature at 2.05 eV. This interpretation of photodestruction experiments for retinal chromophores has in fact a parallel in the theoretical findings[66] and recent experimental reassessment[93] of photodestruction experiments of the photosensitive green fluorescent protein chromophore. We therefore report a range of energies between 2.05 and 2.34 eV as a more conservative experimental estimate of the vertical excitation of retinal chromophores. Our DMC, single-state, and MS-CASPT2 excitations are compatible with the experimental estimate, especially if we consider the remaining uncertainty on the ground-state DFT and MP2 geometries and the fact that we did not include vibrational effects. Setting the IPEA shift to zero moves the vertical CASPT2 excitation toward the lower end of the experimental range, namely, the possible location of the adiabatic transition, and the excitation even falls below the lower bound in the case of the single-state approach. We note that we could not perform CC calculations for the 11-*cis* model with the available codes and that the best CC2 result of 2.10 eV found in the literature[40] is about 0.20 eV lower than the CASPT2 excitation we compute on a similar B3LYP geometry. This discrepancy is rather puzzling since the CASPT2 and CC2 excitation energies agree rather well for all smaller models and could be due to the particular basis used in ref 40 or to the different response of CC2 and CASPT2 to the addition of the $\beta$-ionone ring missing in the smaller models.

## 6. In-Plane Geometrical Optimization

We optimize the in-plane excited-state geometries of the retinal chromophore models (A, B, C, D) using the CASSCF, MS-CASPT2, CC2, CCSD, and VMC approaches. We always follow the second root in the optimization and use two roots in the SA-CASSCF and MS-CASPT2 calculations as well as in the optimization of the VMC wave functions. The CAS expansion correlates all $\pi$ electrons and an equal number of orbitals with the exception of models A and B, where we include more virtual orbitals to be consistent with previous calculations.[15] As shown in ref 22 for model A, a smaller active space of 6 electrons in 6 orbitals yields equivalent CASSCF results. We impose the planarity of the conjugated chain by constraining the optimization to $C_s$ symmetry, and unless otherwise stated, we start the excited-state optimization from the DFT/B3LYP ground-state geometry.

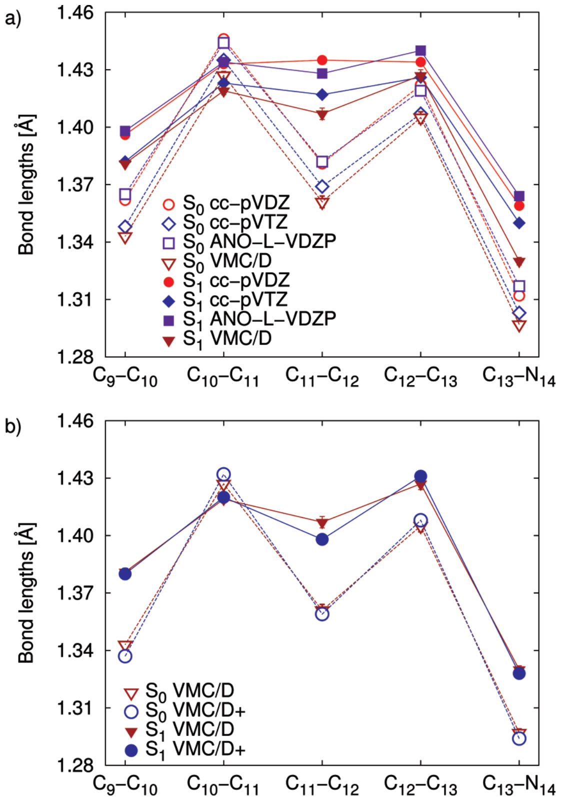


***Figure 4.*** Bond lengths (Å) of the PSB3(0) model (A) optimized in the ground and excited states with the CASPT2 (panel a) and VMC (panel b) approaches and different basis sets. The CASPT2 geometries are computed with the cc-pVDZ, cc-pVTZ, and ANO-L-VDZP basis, and the VMC results with the D and D+ basis sets. In panel a, the VMC/D bond lengths are also shown for comparison. Planar symmetry is imposed.

**6.1. Dependence on Basis Set and Other Parameters.** In all geometrical optimizations, we employ the cc-pVDZ basis set. As shown in Figure 4a for the minimal model (A), the effect of using the larger cc-pVTZ basis set is to systematically shorten all ground- and excited-state CASPT2 bond lengths by about 0.010−0.015 Å without affecting the bond length pattern, as was also previously observed in ref 20. Differently from the case of the excitation energies, the ANO-L-VDZP basis yields comparable bond lengths to the cc-pVDZ value, which only disagree by 0.06 and 0.07 Å in the $C_{11}-C_{12}$ and $C_{12}-C_{13}$ excited-state bonds, respectively. A similar behavior as a function of the size of the basis set is also found for the CASSCF and DFT bond lengths, although the shortening in not as pronounced as for the CASPT2 results. In Figure 4b, we compare the VMC results obtained with the D and D+ basis sets, which are almost equal. Interestingly, the VMC results obtained with the D (cc-pVDZ) basis are very close to the CASPT2/cc-pVTZ results, so the presence of the Jastrow factor appears to compensate for the use of a smaller basis.

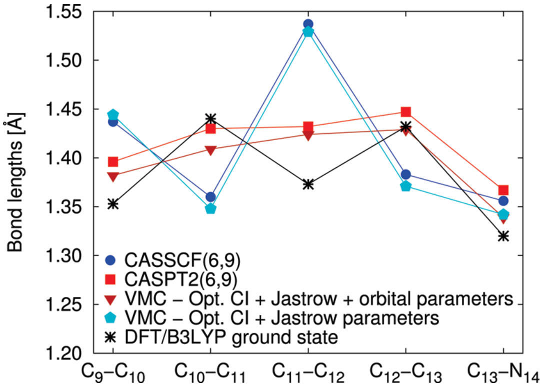


***Figure 5.*** VMC bond lengths (Å) in the excited state of the PSB3(1) model (B) computed with two different wave functions. In one case, only the CI and Jastrow parameters are optimized within energy minimization in a state-average fashion while, in the other, all (orbital included) parameters are optimal. Planar symmetry is imposed.

The VMC geometrical optimization is very sensitive to the quality of the trial wave function as shown for model B in Figure 5. We start the optimization from the Franck−Condon region, and if we optimize only the Jastrow and CI coefficients within VMC, we obtain a VMC minimum which corresponds to the bond-inverted CASSCF geometry. On the other hand, if we include the orbital parameters in the VMC optimization, we obtain a very different VMC geometry which agrees with the minimum obtained by the other highly correlated approaches as shown below. Thus, in the VMC geometrical relaxation, we need to optimize all wave function parameters. We also note that preliminary calculations with DMC gradients indicate that the use of DMC does not mend the behavior of VMC when the DMC gradients are computed from wave functions with only optimal Jastrow and CI parameters.

**6.2. Results.** To understand how the geometry of the retinal chromophore is modified upon photoexcitation, we begin with the minimal model (A) and show in Figure 6a the ground-state bond lengths as obtained with the CASSCF, CASPT2, MP2, VMC, and DFT/B3LYP approaches. All methods agree in predicting a strong single−double bond-length alternation with a short, double bond between the central carbons. The MP2 and CASPT2 geometries are almost exactly equal since the ground state is dominated by a single configuration (89% weight) and CASPT2 is equivalent to MP2 for a single-reference CASSCF wave function. The DFT/B3LYP bond lengths deviate from the MP2 and CASPT2 values by at most 0.01 Å in the two single bonds. The VMC bond lengths are shorter by about 0.015 Å, and this can be explained as a basis set effect as discussed above. Only the CASSCF approach is at variance with the other approaches in the sense that it exhibits a greater bond length alternation, as has also been observed for larger retinal models.[20,26] The difference between CASSCF and the other approaches is on the order of 0.01−0.02 Å for model A but grows as the model becomes larger (see Figure SI-5, Supporting Information). In view of these results, we find that the DFT/B3LYP approach offers a good balance between performance and computational cost for the computation of the ground state structure.

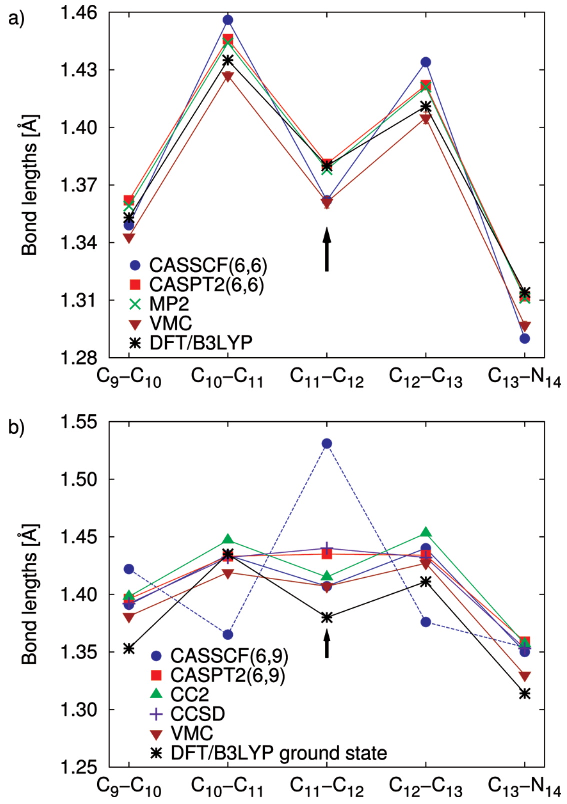


***Figure 6.*** Bond lengths (Å) of the PSB3(0) model (A) optimized in the ground (panel a) and excited (panel b) states with the CC2, CCSD, CASSCF, CASPT2, and VMC methods. The DFT/B3LYP and MP2 ground states are also shown. The cc-pVDZ basis is used and planarity imposed. CASSCF displays two minima in the excited state.

The excited-state bond lengths of the minimal model (A) are shown in Figure 6b. The CASSCF approach exhibits two almost degenerate minima, while all other approaches yield only one minimum. The first CASSCF minimum (solid line) displays a lengthening of almost all bonds and a largely preserved bond-length pattern as compared to the ground state. The second CASSCF minimum (dashed line) is about 0.022 eV higher in energy than the other CASSCF minimum and displays a pronounced bond-length inversion with respect to the ground state. Importantly, we note that the first minimum is found when starting the optimization from the ground-state geometry, while we started from a geometry biased toward bond inversion to find the second one. In addition, regardless of the starting point, we only converge to a single CASSCF minimum, corresponding to the first minimum, if the ANO-L-VDZP basis set is used instead of the cc-pVDZ basis set. For the two CASSCF minima obtained with the cc-pVDZ basis, we report the wave function character and orbitals in the Supporting Information.

As for the other methods, we observe that most bond lengths become longer and more similar in the excited state. The CC2 and VMC structures largely preserve the short−long bond-length pattern of the ground state as observed for the first CASSCF minimum, while CASPT2 and CCSD give three middle bonds of almost equal length. At the CASPT2 level, we also investigated extensively the existence of a bond-inverted minimum by starting the excited-state optimization from geometries biased toward bond inversion but

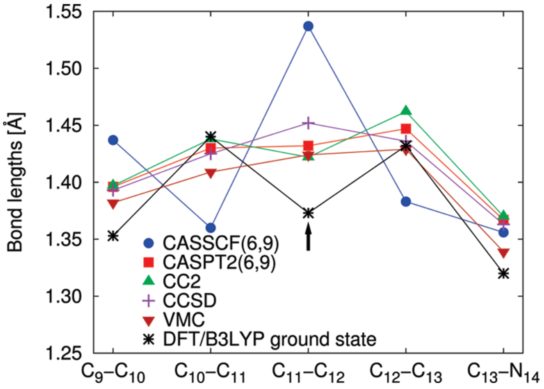


***Figure 7.*** Bond lengths (Å) of the PSB3(1) model (B) optimized in the excited state with the CC2, CCSD, CASSCF, and CASPT2 methods. The DFT/B3LYP ground-state bond lengths are also shown. The cc-pVDZ basis is used and planar symmetry imposed. Differently from model A without the methyl (Figure 6), CASSCF only displays here one minimum.

could not locate a second minimum. Our CASSCF and CASPT2 results are consistent with the study by Page and Olivucci[20] using the 6-31G(d) basis set.

Surprisingly, adding a methyl group to the minimal model (A) to generate model B has profound effects on the bond lengths, as shown in Figure 7. In particular, there is now only one CASSCF minimum which exhibits a pronounced bond length inversion as compared to the ground state and is at variance with the results obtained with all other approaches. The differences among the results obtained with the other methods is instead significantly smaller. The CC2 geometry of model B is similar to model A with a lengthening of most bonds and a largely preserved bond-length pattern with respect to the ground state. Similarly to model A, CASPT2 yields close to equal bond lengths for the three middle bonds, with the $C_{12}-C_{13}$ bond being the longest, while CCSD gives the middle $C_{11}-C_{12}$ bond as being slightly larger. The VMC minimum displays a similar bond length pattern as CASPT2 but shorter absolute bond lengths, which can be explained as a basis set effect as explained above.

When going to larger models, we find that CASSCF yields only one minimum where the short–long bond-length pattern is inverted with respect to the ground state as in the case of model B. In Figure 8, we show the excited-state bond lengths for model C and observe that the CASSCF minimum with bond-length inversion is at variance with all other approaches. The CASPT2 and CC2 are very close to each other and exhibit a largely preserved bond length pattern and overall lengthening of most bonds with respect to the ground state. The CCSD geometry displays no distinct bond-length pattern and an overall lengthening of most bonds, and the VMC gives a similar bond-length pattern as CCSD but shorter bond lengths as seen above.

In Figure 9, we show the excited-state bond lengths of model D, which has the full conjugated chain of the retinal chromophore and only misses the $\beta$-ionone ring. For this model, we only show the bond lengths obtained with the CASSCF and CASPT2 approaches. In addition, we also show CASPT2 results obtained with the ANO-L-VDZP basis set. As for models B and C, CASSCF gives a structure characterized by bond-length inversion with respect to the ground state. However, it is now the CASPT2 approach which gives two profoundly different minima. The first CASPT2 minimum (solid line) is similar to the CASPT2 geometry of model C with a preserved bond length pattern and overall lengthening of most bonds as compared to the ground state. The second CASPT2 minimum (dashed line) is very close to the CASSCF geometry and is about 0.045 eV higher in energy than the first CASPT2 minimum. Importantly, we note that the first CASPT2 minimum is found when starting from the ground-state geometry, while the second CASPT2 minimum is reached when starting from the CASSCF excited-state geometry. Moreover, the existence of this second minimum is dependent on the choice of the basis: The two CASPT2 minima obtained with the cc-pVDZ basis are also found when the 6-31G(d) basis set is used (not shown in the figure), while only the first CASPT2 minimum with no bond length inversion is obtained regardless of the starting geometry when the ANO-L-VDZP basis

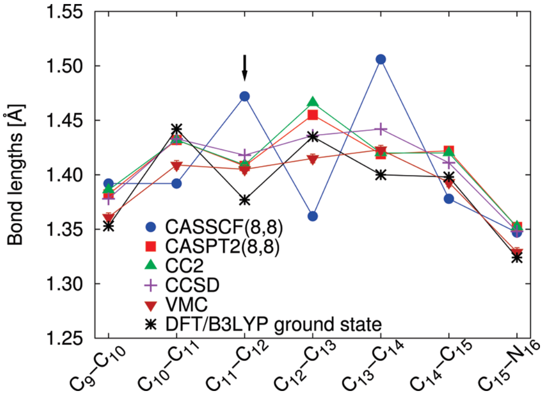


***Figure 8.*** Bond lengths (Å) of the PSB4(1) model (C) optimized in the excited state with the CC2, CCSD, CASSCF, CASPT2, and VMC methods. The DFT/B3LYP ground-state bond lengths are also shown. The cc-pVDZ basis is used and planar symmetry imposed.

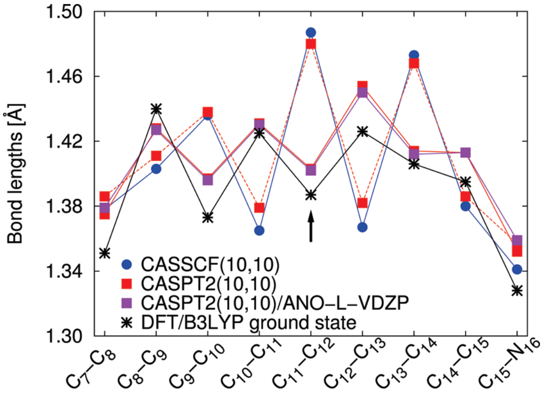


***Figure 9.*** Bond lengths (Å) of the PSB5(1) model (D) optimized in the excited state with the CASSCF and CASPT2 methods and the cc-pVDZ basis. We also show the CASPT2/ANO-L-VDZP results. CASPT2 displays two minima in the excited state with the cc-pVDZ basis and only one minimum with the ANO-L-VDZP basis. The DFT/B3LYP ground-state bond lengths are also shown. Planar symmetry is imposed.

set is used. These results seem to indicate that the bond-inverted CASPT2 structure is a spurious local minimum with no chemical significance, which is not reached when the optimization is started from the ground-state structure, that is, upon photoexcitation. We finally observe that a previous CASSCF and CASPT2 study by Page and Olivucci[20] using the 6-31G(d) basis set reports an excited-state CASPT2 structure of model D characterized by bond inversion. This finding can be easily explained by the fact that they started the CASPT2 geometrical optimization from the excited-state CASSCF minimum and were thus not able to reach the other minimum.

In summary, we see that the CASSCF excited-state geometries are at variance with the CASPT2, CC, and QMC results with the exception of the minimal model (A) where CASSCF displays two minimum structures, one of which is in agreement with the geometries obtained by the other approaches. The minimal model appears however to be a special case since the addition of a single methyl group in model B changes the picture and breaks the agreement between CASSCF and the other approaches. The inadequacy of CASSCF in describing in-plane excited structures of the retinal chromophore is also apparent from the results obtained for all the larger models.

## 7. Out-of-Plane Relaxation

**7.1. Minimum Energy Paths.** We determine the excited-state MEP of the retinal models B and C using the CASSCF and CASPT2 approaches. Ground- and excited-state CASSCF MEPs have previously been calculated for several retinal models,[15,16,18,22,33] and the common assumption is that the effect of dynamical correlation can be in part recovered by simply computing the CASPT2 energy on the final CASSCF geometries (CASPT2//CASSCF). Our aim here is to assess the validity of this assumption for the retinal chromophores by comparing the CASSCF and CASPT2 MEPs. To the best of our knowledge, the CASPT2 method has not been used to determine MEPs for the retinal models since CASPT2 energy gradients are substantially more expensive than CASSCF ones and still considered too costly for the routine investigation of these systems.[38,55] In the literature, we only found a CASPT2 study performing a constrained excited-state potential energy surface scan for the minimal model (A).[41]

The MEP calculations are performed using the steepest descent path optimization scheme implemented in MOLCAS 7.2 and described in ref 94. The procedure consists of a series of constrained geometrical optimizations in mass-weighted coordinates and yields the intrinsic reaction path. In each optimization, the potential energy is minimized on a hypersphere of a chosen radius, centered at a given reference structure. The CASSCF and CASPT2 ground-state geometries define the Franck−Condon point and initial reference structure for the corresponding MEP calculations. The radius of the hypersphere is either 0.06 or 0.1 au for model B and 0.1 au for model C. Upon convergence of the constrained geometrical optimization, the obtained minimum structure on the hypersphere is taken as new reference structure, and the procedure is iterated. As in the planar optimizations, the state averaging in the CASSCF and CASPT2 includes only the ground ($S_0$) and first excited state ($S_1$) since the next state is significantly higher in energy and does not play an active role (see Supporting Information).

We define the torsional angle $\theta$ as the $C_{10}-C_{11}-C_{12}-C_{13}$ dihedral angle and the torsional angle $\gamma = 180° - \phi$ where $\phi$ is the $C_{11}-C_{12}-C_{13}-C_{14}$ dihedral angle and $\gamma$ is taken in the range from $-180°$ to $+180°$. Both torsional angles have a value of 0° in the ground state and indicate the deviation for planarity. These angles correspond to the torsional motion around the $C_{11}-C_{12}$ and $C_{12}-C_{13}$ bonds which are double and single in the ground state, respectively. We note that geometries corresponding to the angles $(\theta, \gamma)$ and $(-\theta, -\gamma)$ are equivalent since the molecules are planar in the ground state and there is no preferential direction for torsion.

In Figure 10, we show the results from the MEP calculation for model B and report the energies, the bond lengths for the formal double and single bonds along the conjugated chain, and the torsional angle $\theta$ for the central $C_{11}-C_{12}$ *cis* bond. The CASSCF MEP is characterized by two sequential modes. The initial relaxation is toward a planar structure similar to the CASSCF $C_s$ minimum discussed above, which exhibits bond-length inversion with respect to the ground state, with the central $C_{11}-C_{12}$ bond being the longest in the excited state. This in-plane motion is followed by a torsion around the central bond toward an angle $\theta$ of about 65°, where a conical intersection region is encountered and the excited-state MEP is stopped. The CASSCF MEP is barrierless, while there is a small barrier of about 0.1 eV in the CASPT2//CASSCF energies.

The CASPT2 MEP is distinctly different from the CASSCF one even though the final outcome of the photoisomerization process is similar. The first difference is that the initial planar relaxation is toward a structure similar to the CASPT2 $C_s$ minimum, which is therefore not characterized by bond inversion. The three middle bonds become almost equal, and the $C_{12}-C_{13}$ bond, which is long in the ground state, is the longest in the excited state. The subsequent torsional motion is around the central $C_{11}-C_{12}$ bond where we observe a plateau in the excited-state energy up to an angle $\theta$ of about 22°. When $\theta$ is about 17°, the three middle bonds begin to change dramatically: The central $C_{11}-C_{12}$ bond lengthens while the two neighboring bonds shorten, so their lengths become similar to those of the CASSCF MEP. The excited-state energy starts then decreasing at a faster pace, and the torsional motion continues toward $\theta \approx 69°$ where a conical intersection region is encountered and the excited-state MEP is stopped. A similar behavior is observed in the constraint excited-state optimization of the minimal model (A) in ref 41, where an energy plateau is observed for $\theta$ between 0° and 25°, followed by a sudden drop in the energy and change in geometry between 25° and 30°. In addition, studies on the minimal model (A) have found that conical intersection geometries obtained with CASSCF and CASPT2 are very similar.[19,20,23] This is consistent with the results obtained here as the CASSCF and CASPT2 MEPs show similar structures near the conical intersection.

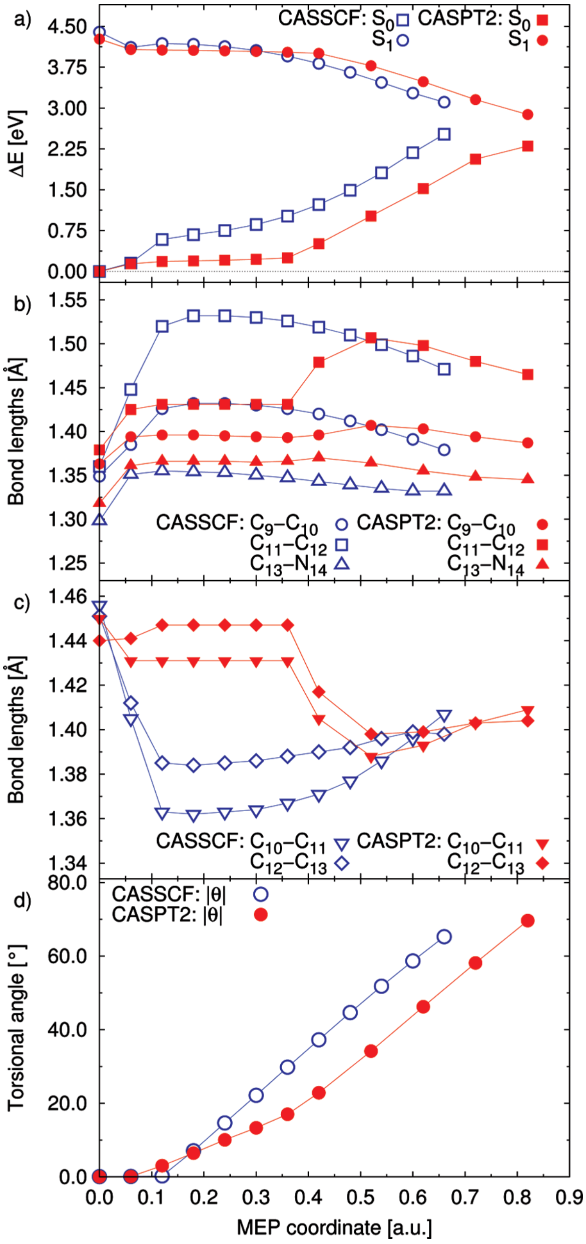


***Figure 10.*** CASSCF and CASPT2 excited-state MEPs for the PSB3(1) model (B), obtained with a CAS(6,9) expansion and the cc-pVDZ basis. We report the CASPT2//CASSCF and CASPT2//CASPT2 ground- and excited-state energies (a), the bond lengths for formal double (b) and single bonds (c), and the absolute value of the torsional angle $\theta$ around the central $C_{11}-C_{12}$ bond (d). All energies are relative to the ground-state energies of the CASSCF and CASPT2 ground-state geometries, which are the starting point of the corresponding MEPs.

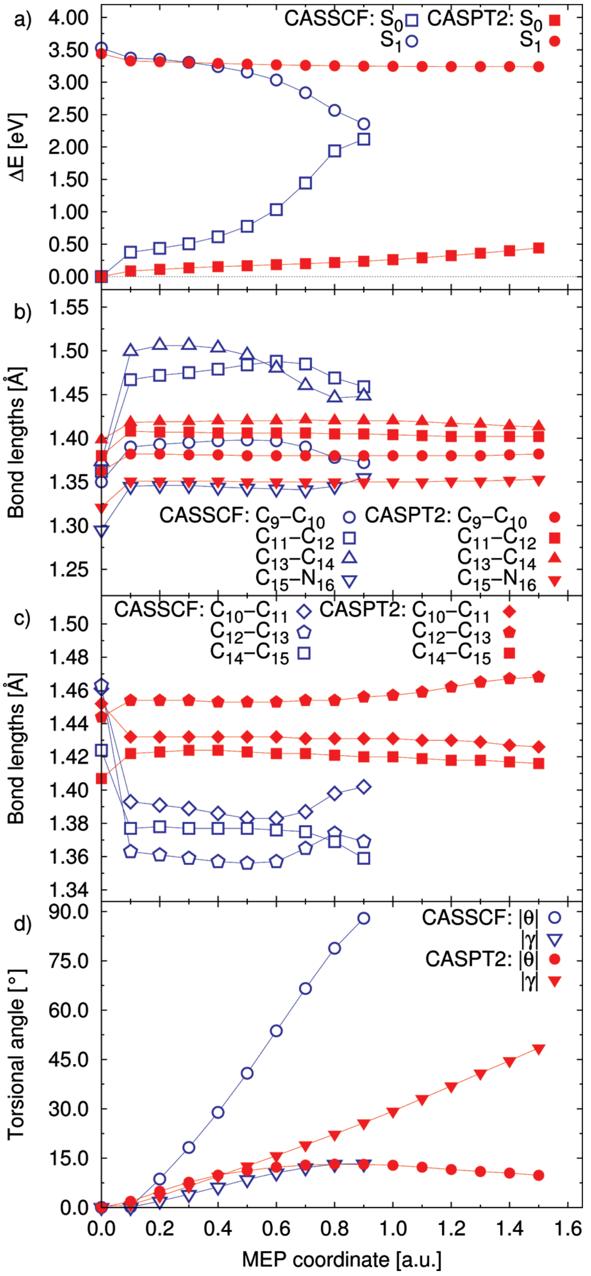


***Figure 11.*** CASSCF and CASPT2 excited-state MEPs for the PSB4(1) model (C), obtained with a CAS(8,8) expansion and the cc-pVDZ basis. We report the CASPT2//CASSCF and CASPT2//CASPT2 ground- and excited-state energies (a), the bond lengths for formal double (b) and single bonds (c), and the absolute values of the torsional angles $\theta$ and $\gamma$ around the $C_{11}-C_{12}$ and the $C_{12}-C_{13}$ bonds, respectively (d). For CASSCF, $\theta < 0°$ and $\gamma < 0°$, while, for CASPT2, $\theta > 0°$ and $\gamma < 0°$. All energies are relative to the ground-state energies at the CASSCF and CASPT2 ground-state geometries, which are the starting point for the corresponding MEPs.

To investigate the effect of lengthening the conjugated chain, we compute the MEP of model C, as shown in Figure 11. The CASSCF and CASPT2 approaches give a different isomerization mechanism, and the relevant torsional angles are not only $\theta$ around the $C_{11}-C_{12}$ bond (formal double) but also $\gamma$ around the $C_{12}-C_{13}$ bond (formal single). The CASSCF MEP is similar to the one of model B and is characterized by two sequential modes, namely, an initial in-plane bond-length inversion followed by a torsional motion around the $C_{11}-C_{12}$ bond until the conical intersection region is encountered at $\theta \approx 88°$. There is also a small torsion around the $C_{12}-C_{13}$ bond with an angle $\gamma \approx 13°$ at the end of the MEP. Differently from model B, the CASPT2//CASSCF excited-state energies show no barrier.

The CASPT2 MEP is rather different from the CASSCF one. The initial relaxation is toward a planar structure which is similar to the CASPT2 $C_s$ minimum and exhibits a largely preserved bond-length pattern with respect to the ground state and an overall lengthening of most bonds. This in-plane motion is followed by a concerted increase of $\theta$ (also active in the CASSCF isomerization) and $\gamma$ up to a MEP coordinate of 0.5 au. Beyond this point, $\gamma$ keeps increasing while $\theta$ changes only slightly, so the molecule is twisting only around the $C_{12}-C_{13}$ bond (formal single), while all bond lengths remain almost constant. At a MEP coordinate of 1.5 au ($\gamma \approx -49°$), a barrier is encountered and the MEP optimization cannot proceed further. Both the ground- and excited-state energies vary very little along the whole MEP, and both states display a long plateau. At the final MEP coordinate, the excited-state energy is only 0.20 eV lower than the Franck−Condon point and the ground-state energy higher by about 0.44 eV, so the vertical excitation has decreased from 3.44 to 2.80 eV.

In order to compare the CASSCF and CASPT2 isomerization mechanisms with the CC2 results, we also perform straight geometrical excited-state optimization with all three approaches since the code we use to perform CC2 calculations does not have the capability of computing MEP. For consistency, all optimizations are started from the DFT/B3LYP ground-state geometries. For model B, all the approaches yield isomerization around the central $C_{11}-C_{12}$ bond and proceed toward the same final point in the conical intersection region. However, from the CASSCF and CASPT2 MEP results, we know that the isomerization proceeds rather differently even though the final structures are equivalent. Therefore, we cannot infer too much about the behavior of CC2 from the agreement of the method on the final structure of model B but proceed with model C, where the final outcomes of the CASSCF and CASPT2 MEP are distinctively different.

We show the optimal CC2 and CASPT2 excited-state structures of model C in Figure 12. We observe that CC2 isomerizes around the $C_{12}-C_{13}$ bond as CASPT2, while CASSCF is consistent with the MEP behavior and yields isomerization around the $C_{11}-C_{12}$ bond (not shown in the figure). The CASPT2 optimal geometry has a torsional angle $\gamma = 43.6°$ and is energetically between the MEP geometries at 1.3 and 1.4 au. Even though the isomerization is around the same bond, the optimal CC2 torsional angle of $\gamma = 100.1°$ is however significantly different from the CASPT2 value. To understand this difference, we investigate the possible existence of a barrier in the CASPT2 potential energy surface and perform a constrained excited-state geometrical optimization in CASPT2 by varying the angle $\gamma$ between 45° and 85°. The resulting excited-state energies are shown in Figure 13 and display a small barrier of about 0.03 eV. If we perform an excited-state CASPT2 optimization starting from the constrained structure just beyond the barrier, we recover a minimum excited-state structure which has a torsional angle of $\gamma = 112.7°$ (Figure 12c) and is in much closer agreement with the CC2 optimal geometry. The CASPT2 excited-state energy is only 0.09 eV lower than the value for the minimal structure at $\gamma = 43.6°$. An analysis of the CASPT2 geometries along the constrained path of Figure 13 reveals that the origin of the barrier is due to steric interactions of the methyl group with the nearby hydrogens since the main difference between the geometries before and after the barrier is a small rotation of the methyl group. We also note that a previous CC2 investigation on model C without the methyl group [PSB4(0)] found a small barrier of 0.01 eV at $\gamma \approx 30°$ and an absolute minimum at about 100°.[29] Therefore, the apparent presence/absence of a barrier in the CASPT2/CC2 optimization may possibly be due to the particular geometrical optimization algorithm used in the different codes or to slightly different initial configurations in the optimization procedure.

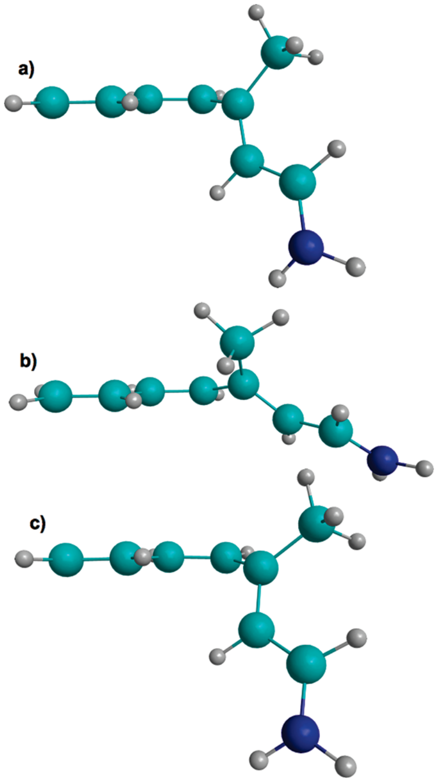


***Figure 12.*** CC2 (a) and CASPT2 (b) excited-state optimal structures of the PSB4(1) model (C), obtained by starting the optimization from the DFT/B3LYP ground-state geometry. The CC2 torsional angles are $\gamma = 100.1°$ and $\theta = 2.6°$, while CASPT2 yields $\gamma = 43.6°$ and $\theta = -10.5°$. The CASPT2 structure (c) is obtained by starting the optimization from the constrained structure just beyond the barrier ($\gamma = 75°$) in Figure 13 and has angles $\gamma = 112.7°$ and $\theta = 8.1°$.

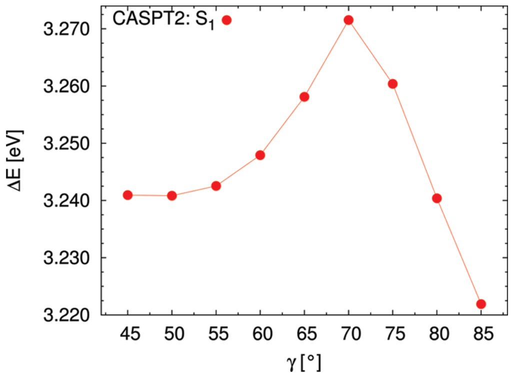


***Figure 13.*** CASPT2 excited-state energies of the PSB4(1) model (C) optimized at constrained torsional angles, $\gamma$, from 45° to 85°. The energy is shown relative to the ground-state value at the starting point of the CASPT2 MEP (Figure 11). A CAS(8,8) expansion and the cc-pVDZ basis set are used.

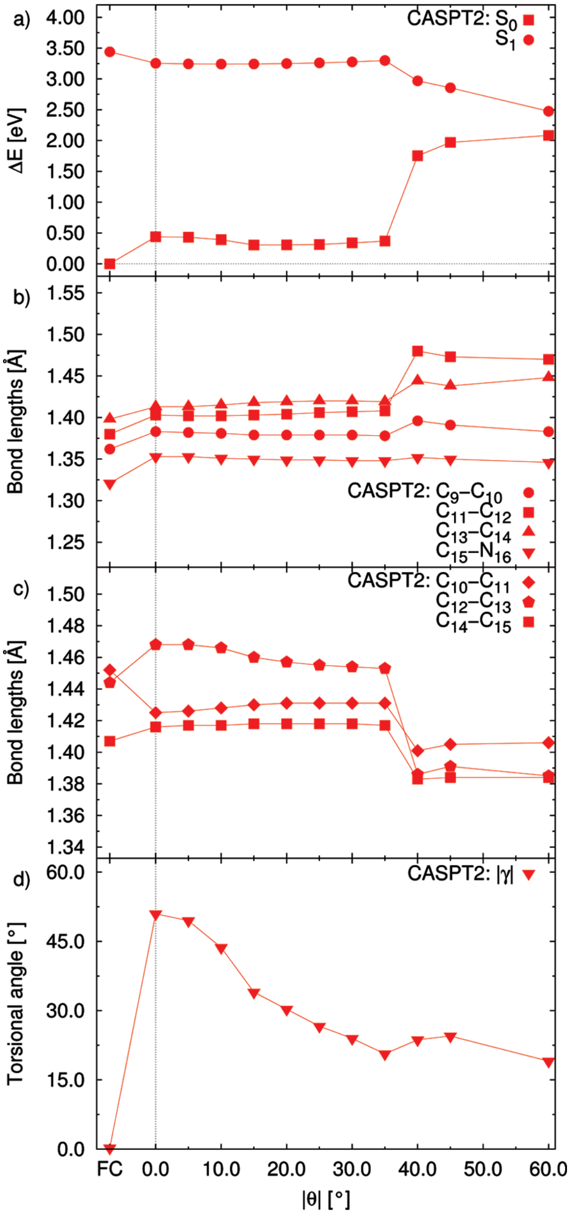


***Figure 14.*** CASPT2 excited-state optimization of the PSB4(1) model (C) at constrained torsional angles, $\theta$, from 0° to 60°. We report the CASPT2//CASPT2 ground- and excited-state energies (a), the bond lengths for formal double (b) and single bonds (c), and the absolute value of the torsional angle $\gamma$ around the $C_{12}-C_{13}$ bond (d). The quantities computed at the CASPT2 Franck−Condon (FC) point are also shown in all panels. A CAS(8,8) expansion and the cc-pVDZ basis set are used. For the torsional angle, $\theta < 0°$ and $\gamma > 0°$.

**7.2. Reactive versus Nonreactive Paths.** The CASPT2 MEP of the retinal model C gives isomerization around a single bond, does not lead to a conical intersection region, and corresponds to a nonreactive path. To investigate whether a rotation around a double bond may give a reactive path and lead to a photoproduct, we optimize the excited-state CASPT2 geometry of model C at constrained torsional angles, $\theta$, around the $C_{11}-C_{12}$ *cis* bond and show the results in Figure 14.

At $\theta = 0°$, the molecule is unstable toward single-bond rotation, which is not surprising since the CASPT2 MEP gives isomerization around the same single bond and is always characterized by small values of the angle $\theta$ (less than 10°). The resulting constrained geometry has an angle $\gamma$ of about 51° and is in fact very similar to the last point of the CASPT2 MEP. If we increase $\theta$ from 0° to 35°, the angle $\gamma$ diminishes while the bond lengths become closer to the values in the initial part of the CASPT2 MEP. Concurrently, the excited-state energy rises and displays a small barrier of about 0.06 eV, which peaks at $\theta = 35°$. The barrier is overcome at $\theta = 40°$, where we suddenly observe bond inversion and a large increase in the ground-state energy and a decrease in the excited-state energy. The degree of bond inversion is however not as pronounced as in the CASSCF MEP, and the geometries are characterized by a larger residual rotation around the single bond. If we further increase $\theta$, the excitation energy continues to decrease, and we encounter a conical intersection region. The CASPT2 isomerization around the double bond corresponds therefore to a reactive path which is characterized by a small barier and eventually leads to a conical intersection region whose topology is rather similar to the CASSCF one.

To assess the behavior of the CC approach, we also perform constrained CC2 optimization around the double bond. The CC2 optimization at small values of $\theta$ leads to a single-bond rotation with very large values of $\gamma$ (greater than 90°). This is compatible with the previous observation that the small steric barrier observed in CASPT2 (see Figure 13) is practically absent in the single-bond isomerization at the CC2 level. If we increase $\theta$ up to 60° and always start the optimization from the optimal constrained geometry at the previous angle, we cannot sufficiently reverse the large rotation around the single bond and the excited-state energy increases instead of decreasing. To assess the existence of a path leading to a conical intersection, we follow therefore a different procedure and simply compute the CC2 energies on the optimal constrained CASPT2 geometries of Figure 14. We find that the ground- and excited-state CC2 energies are in very good agreement with the CASPT2 values up to $\theta = 45°$. As expected and also discussed in ref 37, CC2 encounters convergence problems at larger values of $\theta$ as the system is approaching the conical intersection region. Consequently, the use of CC2 confirms the existence of a reactive path which corresponds to double-bond rotation, displays a small barrier, and leads to lower excited-state energies. However, the approach is not suitable for following the system through the conical intersection toward a photoproduct.

## 8. Discussion and Conclusions

We have presented a systematic investigation of model retinal chromophores in the gas phase with special emphasis on geometrical relaxation in the excited state. One aim of the work is to assess the relative performance of very diverse computational approaches as CASSCF, CASPT2, CC, and QMC in describing conformational changes in the excited states. The other major goal is to determine the validity of the generally accepted picture resulting from CASSCF calculations that the excited-state relaxation of retinal chromophores proceeds via bond inversion and torsional motion around formal double bonds. Differently from previous studies, we employ approaches such as CASPT2 and QMC, which are superior to CASSCF as they offer a balanced description of both dynamical and static correlations.

We have also computed the vertical excitations of the retinal models using CASPT2, CC, and QMC, and we begin our discussion with a few comments on these results. We find that the CC and DMC methods give similar excitations for all retinal models and that the CASPT2 excitations are quite sensitive to the internal parameters of the theory. In particular, the excitations computed with the IPEA zero-order Hamiltonian are in close agreement with the CC and DMC values, while resorting to the original CASPT2 formulation lowers the excitations by as much as 0.3 eV. The IPEA Hamiltonian was developed to give on average more accurate excitations,[79] and its use is here corroborated by the good agreement with other highly correlated approaches. We also find that the IPEA excitations are more robust as they converge faster with the size of the CAS expansion and are less sensitive to the use of a single- or multistate approach.

For a comparison with experiments, we consider the 11-*cis* chromophore where gas phase photodestruction spectroscopy experiments are available.[92] To interpret the complex absorption spectrum of retinal chromophores, we follow the recent reassessment of similar experiments on a different chromophore[93] and suggest that the lowest-energy peak may correspond to the adiabatic transition while the vertical lies in the broad shoulder around 2.34 eV. Our CASPT2 and DMC vertical excitations computed on the DFT and MP2 ground-state geometries span an energy range of 2.2−2.4 eV, which is consistent with this experimental estimate especially given that we did not include vibrational effects which are strong in this system. The excitations computed on the CASSCF geometry are instead significantly higher but can be discarded as our CASPT2 optimizations of planar retinal models show that DFT and MP2 give more accurate geometries than the CASSCF approach.

We discuss now the core of our work and analyze the performance of the various theoretical approaches in describing the excited-state relaxation of retinal models. Our in-plane optimization of the retinal chromophores indicates that the excited-state structures optimized with CASPT2, CC2, CCSD, and VMC agree rather closely, while they are at variance with the CASSCF geometries. The CASSCF approach gives strong bond inversion in the excited state, which is not observed when optimizing the structures with the other approaches. According to CASPT2, CC, and VMC, photoexcitation weakens all bonds, which stretch and become partly more similar in length while preserving the general bond-length pattern of the ground state. To investigate a nontrivial out-of-plane relaxation, we need to consider a chromophore larger than the model with three double bonds (A or B) since we find that model B isomerizes around the central bond at both the CASPT2 and CASSCF levels, even though the initial skeletal relaxation proceeds rather differently in the two approaches. Therefore, we investigate the minimal energy path for the out-of-plane motion of model C with four double bonds and find that excited-state relaxation at the CASPT2 level proceeds preferentially via a torsional motion around a bond which is formally single in the ground state in agreement with the previous CC calculations by Send and Sundholm.[29,31,35,37] This torsional motion stops at an angle of about 45° and does not lead to a conical intersection region. On the other hand, in the CASSCF approach, bond inversion is followed by torsion around the *cis* bond, and the molecule is immediately funneled into a conical intersection region from where isomerization can proceed toward the *trans* product. To investigate the existence of a reactive path at the CASPT2 level, we also consider the constrained excited-state optimization of model C around the *cis* double bond and find a small barrier to isomerization at rather large angles of rotation. Beyond this barrier, the model finally reaches the conical intersection region similarly to the CASSCF approach.

In summary, our CASPT2 results support the picture of a very flexible retinal chromophore in the excited state, where photoexcitation lengthens all bonds so that torsional motion around nearly any bond may contribute to the dynamics. These findings are consistent with recent CC studies[37] which show that retinal models in the excited state have small or vanishing torsional barriers around both formal single and double bonds. This picture must be contrasted to the results of CASSCF calculations, which give a stiff chromophore that can only twist around formal double bonds. The flexibility of the excited chromophore in the gas phase observed in CASPT2 and CC calculations is also compatible with the observation in solution experiments of the existence of multiple minima possibly corresponding to different torsional conformations.[67] Moreover, it has been proposed that the multiexponential decays observed in solution are related to the possible presence of multiple excited-state paths, some of which are reactive and lead to the photoproduct via the crossing of a conical intersection region, while others are nonreactive, do not lead to conical intersection, and are dominant in solution.[67] This interpretation is compatible with our observation of torsional motion around formal single bonds, which is favored starting from the Franck−Condon region, stops at intermediate angles, and does not lead to photoproducts via a conical intersection.

Finally, our results demonstrate the importance of including a balanced description of dynamical and static correlation in the computation of the excited-state gradients. The favorable comparison with the CASPT2 approach indicates that the CC2 method is a useful tool for the study of retinal systems (at least far from the conical intersection region) and that QMC can give accurate gradients when all parameters in the wave function are optimized in energy minimization. Our results raise serious concerns about the common use of the CASSCF approach to investigate the geometrical relaxation of retinal systems and show that computing single-point CASPT2 excitations on CASSCF geometries to partially include the neglected dynamical correlation is generally not a valid procedure to obtain reliable potential energy surfaces. In conclusion, our findings call for a reinvestigation of the photoisomerization mechanism of retinal chromophores in the gas phase as well as in the protein environment with higher-level methods than the CASSCF approach.

**Acknowledgment.** We thank Massimo Olivucci, Ivano Tavernelli, Robert Send, and Saverio Moroni for useful discussions. We acknowledge the support from the Stichting Nationale Computerfaciliteiten (NCF-NWO) for the use of the SARA supercomputer facilities.

**Supporting Information Available:** SA-CASSCF and SS-CASPT2 excitations of all models; basis set convergence of the SA-CASSCF and CC energies of the PSB3(0) model (A); character of the excited-state CASSCF wave function and orbitals computed at the ground-state B3LYP, excited-state CASPT2, and two excited-state CASSCF planar minima of the PSB3(0) model (A); ground-state bond lengths of the PSB5(1) model (C); CASPT2 energies computed along the out-of-plane excited-state relaxation paths of models PSB3(1) (A) and PSB4(1) (B) and obtained with a state average over three states in the CASSCF and CASPT2 calculations; ground- and excited-state coordinates of all retinal models. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Wald, G. *Science* **1968**, *162*, 230–239.

(2) Okada, T.; Sugihara, M.; Bondar, A.-N.; Elstner, M.; Entel, P.; Buss, V. *J. Mol. Biol.* **2004**, *342*, 571–583.

(3) Kandori, H.; Shichida, Y.; Yoshizawa, T. *Biochemistry (Moscow)* **2001**, *66*, 1483–1498.

(4) Schoenlein, R. W.; Peteanu, L. A.; Mathies, R. A.; Shank, C. V. *Science* **1991**, *254*, 412–415.

(5) Kandori, H.; Katsuta, Y.; Ito, M.; Sasabe, H. *J. Am. Chem. Soc.* **1995**, *117*, 2669–2670.

(6) Becker, R. S.; Freedman, K. *J. Am. Chem. Soc.* **1985**, *107*, 1477–1485.

(7) Kim, J. E.; Tauber, M. J.; Mathies, R. A. *Biochemistry* **2001**, *40*, 13774–13778.

(8) Kandori, H.; Furutani, Y.; Nishimura, S.; Shichida, Y.; Chosrowjan, H.; Shibata, Y.; Mataga, N. *Chem. Phys. Lett.* **2001**, *334*, 271–276.

(9) Kobayashi, T.; Saito, T.; Ohtani, H. *Nature* **2001**, *414*, 531–534.

(10) Herbst, J.; Heyne, K.; Diller, R. *Science* **2002**, *297*, 822–825.

(11) McCamant, D. W.; Kukura, P.; Mathies, R. A. *J. Phys. Chem. B* **2005**, *109*, 10449–10457.

(12) Kukura, P.; McCamant, D. W.; Yoon, S.; Wandschneider, D. B.; Mathies, R. A. *Science* **2005**, *310*, 1006–1009.

(13) Kukura, P.; McCamant, D. W.; Mathies, R. A. *Annu. Rev. Phys. Chem.* **2007**, *58*, 461–488.

(14) Kennis, J. T.; Groot, M.-L. *Curr. Opin. Struct. Biol.* **2007**, *17*, 623–630.

(15) Garavelli, M.; Celani, P.; Bernardi, F.; Robb, M. A.; Olivucci, M. *J. Am. Chem. Soc.* **1997**, *119*, 6891–6901.

(16) Garavelli, M.; Vreven, T.; Celani, P.; Bernardi, F.; Robb, M. A.; Olivucci, M. *J. Am. Chem. Soc.* **1998**, *120*, 1285–1288.

(17) Garavelli, M.; Bernardi, F.; Olivucci, M.; Vreven, T.; Klein, S.; Celani, P.; Robb, M. A. *Faraday Discuss.* **1998**, *110*, 51–70.

(18) González-Luque, R.; Garavelli, M.; Bernardi, F.; Merchán, M.; Robb, M. A.; Olivucci, M. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 9379–9384.

(19) De Vico, L.; Page, C. S.; Garavelli, M.; Bernardi, F.; Basosi, R.; Olivucci, M. *J. Am. Chem. Soc.* **2002**, *124*, 4124–4134.

(20) Page, C. S.; Olivucci, M. *J. Comput. Chem.* **2003**, *24*, 298–309.

(21) Wanko, M.; Garavelli, M.; Bernardi, F.; Niehaus, T. A.; Frauenheim, T.; Elstner, M. *J. Chem. Phys.* **2004**, *120*, 1674–1692.

(22) Fantacci, S.; Migani, A.; Olivucci, M. *J. Phys. Chem. A* **2004**, *108*, 1208–1213.

(23) Serrano-Andrés, L.; Merchán, M.; Lindh, R. *J. Chem. Phys.* **2005**, *122*, 104107.

(24) Wanko, M.; Hoffmann, M.; Strodel, P.; Koslowski, A.; Thiel, W.; Neese, F.; Frauenheim, T.; Elstner, M. *J. Phys. Chem. B* **2005**, *109*, 3606–3615.

(25) Tavernelli, I.; Röhrig, U. F.; Rothlisberger, U. *Mol. Phys.* **2005**, *103*, 963–981.

(26) Blomgren, F.; Larsson, S. *J. Comput. Chem.* **2005**, *26*, 738–742.

(27) Cembran, A.; Gonzalez-Luque, R.; Altoe, P.; Merchan, M.; Bernardi, F.; Olivucci, M.; Garavelli, M. *J. Phys. Chem. A* **2005**, *109*, 6597–6605.

(28) Aquino, A. J. A.; Barbatti, M.; Lischka, H. *ChemPhysChem* **2006**, *7*, 2089–2096.

(29) Send, R.; Sundholm, D. *J. Phys. Chem. A* **2007**, *111*, 27–33.

(30) Send, R.; Sundholm, D. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2862–2867.

(31) Send, R.; Sundholm, D. *J. Phys. Chem. A* **2007**, *111*, 8766–8773.

(32) Barbatti, M.; Granucci, G.; Persico, M.; Ruckenbauer, M.; Vazdar, M.; Eckert-Maksić, M.; Lischka, H. *J. Photochem. Photobiol., A* **2007**, *190*, 228–240.

(33) Cembran, A.; González-Luque, R.; Serrano-Andrés, L.; Merchán, M.; Garavelli, M. *Theor. Chem. Acc.* **2007**, *118*, 173−183.

(34) Weingart, O. *J. Am. Chem. Soc.* **2007**, *129*, 10618–10619.

(35) Send, R.; Sundholm, D. *J. Mol. Model.* **2008**, *14*, 717–726.

(36) Szymczak, J. J.; Barbatti, M.; Lischka, H. *J. Chem. Theory Comput.* **2008**, *4*, 1189–1199.

(37) Send, R.; Sundholm, D.; Johansson, M. P.; Pawłowski, F. *J. Chem. Theory Comput.* **2009**, *5*, 2401–2414.

(38) Schapiro, I.; Weingart, O.; Buss, V. *J. Am. Chem. Soc.* **2009**, *131*, 16–17.

(39) Szymczak, J. J.; Barbatti, M.; Lischka, H. *J. Phys. Chem. A* **2009**, *113*, 11907–11918.

(40) Zaari, R. R.; Wong, S. Y. *Chem. Phys. Lett.* **2009**, *469*, 224–228.

(41) Keal, T.; Wanko, M.; Thiel, W. *Theor. Chem. Acc.* **2009**, *123*, 145–156.

(42) Warshel, A. *Nature* **1976**, *260*, 679–683.

(43) Warshel, A.; Barboy, N. *J. Am. Chem. Soc.* **1982**, *104*, 1469–1476.

(44) Liu, R. S.; Asato, A. E. *Proc. Natl. Acad. Sci. U.S.A.* **1985**, *82*, 259–263.

(45) Hayashi, S.; Tajkhorshid, E.; Schulten, K. *Biophys. J.* **2003**, *85*, 1440–1449.

(46) Andruniów, T.; Ferré, N.; Olivucci, M. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 17908–17913.

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

(48) Röhrig, U. F.; Guidoni, L.; Rothlisberger, U. *ChemPhysChem* **2005**, *6*, 1836–1847.

(49) Hoffmann, M.; Wanko, M.; Strodel, P.; Konig, P. H.; Frauenheim, T.; Schulten, K.; Thiel, W.; Tajkhorshid, E.; Elstner, M. *J. Am. Chem. Soc.* **2006**, *128*, 10808–10818.

(50) Coto, P. B.; Strambi, A.; Ferré, N.; Olivucci, M. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 17154–17159.

(51) Sugihara, M.; Hufen, J.; Buss, V. *Biochemistry* **2006**, *45*, 801–810.

(52) Sekharan, S.; Sugihara, M.; Buss, V. *Angew. Chem., Int. Ed.* **2007**, *46*, 269–271.

(53) Bravaya, K.; Bochenkova, A.; Granovsky, A.; Nemukhin, A. *J. Am. Chem. Soc.* **2007**, *129*, 13035–13042.

(54) Fujimoto, K.; Hayashi, S.; Hasegawa, J.-y.; Nakatsuji, H. *J. Chem. Theory Comput.* **2007**, *3*, 605–618.

(55) Frutos, L. M.; Andruniów, T.; Santoro, F.; Ferré, N.; Olivucci, M. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 7764–7769.

(56) Altun, A.; Yokoyama, S.; Morokuma, K. *Photochem. Photobiol.* **2008**, *84*, 845–854.

(57) Altun, A.; Yokoyama, S.; Morokuma, K. *J. Phys. Chem. B* **2008**, *112*, 6814–6827.

(58) Altun, A.; Yokoyama, S.; Morokuma, K. *J. Phys. Chem. B* **2008**, *112*, 16883–16890.

(59) Altun, A.; Yokoyama, S.; Morokuma, K. *J. Phys. Chem. A* **2009**, *113*, 11685–11692.

(60) Tomasello, G.; Olaso-González, G.; Altoeà, P.; Stenta, M.; Serrano-Andreás, L.; Merchaán, M.; Orlandi, G.; Bottoni, A.; Garavelli, M. *J. Am. Chem. Soc.* **2009**, *131*, 5172–5186.

(61) Hayashi, S.; Tajkhorshid, E.; Schulten, K. *Biophys. J.* **2009**, *96*, 403–416.

(62) Schautz, F.; Filippi, C. *J. Chem. Phys.* **2004**, *120*, 10931–10941.

(63) Schautz, F.; Buda, F.; Filippi, C. *J. Chem. Phys.* **2004**, *121*, 5836–5844.

(64) Cordova, F.; Doriol, L. J.; Ipatov, A.; Casida, M. E.; Filippi, C.; Vela, A. *J. Chem. Phys.* **2007**, *127*, 164111.

(65) Tapavicza, E.; Tavernelli, I.; Rothlisberger, U.; Filippi, C.; Casida, M. E. *J. Chem. Phys.* **2008**, *129*, 124108.

(66) Filippi, C.; Zaccheddu, M.; Buda, F. *J. Chem. Theory Comput.* **2009**, *5*, 2074–2087.

(67) Zgrablić, G.; Haacke, S.; Chergui, M. *J. Phys. Chem. B* **2009**, *113*, 4384–4393.

(68) Jensen, F. *Introduction to Computational Chemistry*, 2nd ed.; John Wiley and Sons Ltd: Chichester, U.K., 2007.

(69) Foulkes, W. M. C.; Mitas, L.; Needs, R. J.; Rajagopal, G. *Rev. Mod. Phys.* **2001**, *73*, 33–83.

(70) Nightingale, M. P.; Melik-Alaverdian, V. *Phys. Rev. Lett.* **2001**, *87*, 043401.

(71) Umrigar, C. J.; Toulouse, J.; Filippi, C.; Sorella, S.; Hennig, R. G. *Phys. Rev. Lett.* **2007**, *98*, 110201.

(72) Filippi, C.; Umrigar, C. J. *Phys. Rev. B* **2000**, *61*, R16291–R16294.

(73) Attaccalite, C.; Sorella, S. *Phys. Rev. Lett.* **2008**, *100*, 114501.

(74) We find that varying the cutoff parameter $\varepsilon$ between $10^{-2}$ and 1.0 does not lead to appreciable changes in the root-mean-square fluctuations of the VMC forces ($\sigma$). On the other hand, the use of $\varepsilon = 10^{-3}$ and smaller values results in a significant increase in $\sigma$ for the systems studied in this paper.

(75) Karlström, G.; Lindh, R.; Malmqvist, P.-Å.; Roos, B. O.; Ryde, U.; Veryazov, V.; Widmark, P.-O.; Cossi, M.; Schimmelpfennig, B.; Neogrady, P.; Seijo, L. *Comput. Mater. Sci.* **2003**, *28*, 222–239.

(76) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(77) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K. ; Voth, G. A. ; Salvador, P. ; Dannenberg, J. J. ; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, Revision D.01; Gaussian, Inc.: Wallingford, CT, 2004.

(78) Finley, J.; Malmqvist, P.-Å.; Roos, B. O.; Serrano-Andrés, L. *Chem. Phys. Lett.* **1998**, *288*, 299–306.

(79) Ghigo, G.; Roos, B. O.; Malmqvist, P.-Å. *Chem. Phys. Lett.* **2004**, *396*, 142–149.

(80) Roos, B. O.; Andersson, K. *Chem. Phys. Lett.* **1995**, *245*, 215–223.

(81) Aquilante, F.; Malmqvist, P.-Å.; Pedersen, T. B.; Ghosh, A.; Roos, B. O. *J. Chem. Theory Comput.* **2008**, *4*, 694–702.

(82) *Aces II*, a quantum chemical program package written by Stanton, J. F.; Gauss, J.; Watts, J. D.; Szalay, P. G.; Bartlett, R. J. with contribution from Auer, A. A.; Bernholdt, D. B.; Christiansen, O.; Harding, M. E.; Heckert, M.; Heun, O.; Huber, C.; Jonsson, D.; Juselius, J.; Lauderdale, W. J.; Metzroth, T.; Ruud, K. and the integral packages *MOLECULE* (Almlof, J.; Taylor, P. R.), *Props* (Taylor, P. R.), and *ABACUS* (Helgaker, T.; Jensen, H. A. A.; Jørgensen, P.; Olsen, J.). See also: Stanton, J. F.; Gauss, J.; Watts, J. D.; Lauderdale, W. J.; Bartlett, R. J. *Int. J. Quantum Chem. Symp.* **1992**, *26*, 879. as well as http://www.aces2.de for the current version.

(83) *CFOUR*, a quantum chemical program package written by Stanton, J. F.; Gauss, J.; Harding, M. E.; Szalay, P. G. with contributions from Auer, A. A.; Bartlett, R. J.; Benedikt, U.; Berger, C.; Bernholdt, D. E.; Christiansen, O.; Heckert, M.; Heun, O.; Huber, C.; Jonsson, D.; Juselius, J.; Klein, K.; Lauderdale, W. J.; Matthews, D.; Metzroth, T.; O'Neill, D. P.; Price, D. R.; Prochnow, E.; Ruud, K.; Schiffmann, F.; Stopkowicz, S.; Varner, M. E.; Vázquez, J.; Wang, F.; Watts, J. D. and. the integral packages *MOLECULE* (Almlf, J.; Taylor, P. R.), *PROPS* (Taylor, P. R.), *ABACUS* (Helgaker,

T.; Jensen, H. J.; Jørgensen, P.; Olsen, J.),and *ECP* routines by Mitin, A. V.; van Wlle, C. For the current version, see http://www.cfour.de.

(84) *CHAMP* is a quantum Monte Carlo program package written by Umrigar, C. J.; Filippi, C. and collaborators.

(85) Burkatzki, M.; Filippi, C.; Dolg, M. *J. Chem. Phys.* **2007**, *126*, 234105.

(86) Filippi, C.; Umrigar, C. J. *J. Chem. Phys.* **1996**, *105*, 213–226. As the Jastrow correlation factor, we use the exponential of the sum of three fifth-order polynomials of the electron−nuclear (e−n), the electron−electron (e−e), and the pure three-body mixed e−e and e−n distances. The Jastrow factor is adapted to deal with pseudo-atoms, and the scaling factor $\kappa$ is set to 0.60 a.u.

(87) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M., Jr. *J. Comput. Chem.* **1993**, *14*, 1347–1363.

(88) Casula, M. *Phys. Rev. B* **2006**, *74*, 161102.

(89) Thom, H.; Dunning, J. *J. Chem. Phys.* **1989**, *90*, 1007−1023.

(90) Widmark, P.-O.; Malmqvist, P.-Å.; Roos, B. O. *Theor. Chem. Acc.* **1990**, *77*, 291–306.

(91) We add one s and one p diffuse function on the carbon and the nitrogen using exponents from the aug-cc-pVDZ basis set, taken from EMSL Basis Set Library (http://bse.pnl.gov).

(92) Nielsen, I. B.; Lammich, L.; Andersen, L. H. *Phys. Rev. Lett.* **2006**, *96*, 018304.

(93) Forbes, M. W.; Jockusch, R. A. *J. Am. Chem. Soc.* **2009**, *131*, 17038–17039.

(94) De Vico, L.; Olivucci, M.; Lindh, R. *J. Chem. Theory Comput.* **2005**, *1*, 1029–1037.

CT900692Y

JCTC Journal of Chemical Theory and Computation

# A Complete Thermodynamic Characterization of Electrostatic and Hydrophobic Associations in the Temperature Range 0 to 100 °C from Explicit-Solvent Molecular Dynamics Simulations

Shun Zhu and Adrian H. Elcock*

*Department of Biochemistry, University of Iowa, Iowa City, Iowa 52242*

Received February 4, 2010

**Abstract:** The electrostatic and hydrophobic interactions that dominate the behavior of proteins and other biomolecules exhibit fundamentally different thermodynamic characteristics, and the correct reproduction of these differences is likely to be an important requirement for models that aim to predict the thermodynamics of protein stability and protein−protein interactions. To assess the abilities of some current models to capture these differences, we report here the results of molecular dynamics (MD) simulations examining the association of acetate−methylammonium and methane−methane pairs at 11 different temperatures from −12.5 to 112.5 °C. Simulations were performed using two popular water models (TIP3P and TIP5P), with a total simulation time of 22 $\mu$s. With both water models, we find that the acetate−methylammonium salt-bridge interaction is significantly more stabilized by high temperatures (e.g., over the range 25 to 100 °C) than is the methane−methane hydrophobic interaction. At low temperatures however, the two models exhibit quite different behavior, with the TIP5P model predicting little change in the relative stabilities of the two types of interaction in the range −12.5 to 50 °C; this surprising result has potential implications for understanding adaptation to life in psychrophilic organisms. Fitting the $\Delta G$ data to the Gibbs−Helmholtz equation allows the $\Delta H$, $\Delta S$, and $\Delta C_p$ of interaction to be obtained, thereby yielding a complete thermodynamic characterization of the different types of interaction in the temperature range 0 to 100 °C: despite significant quantitative differences, both water models correctly capture the opposite signs of the $\Delta C_p$ of electrostatic and hydrophobic interactions. Finally, we show that at high temperatures a Poisson-based continuum solvation model provides good agreement with the explicit-solvent MD results, but only when the atomic radii used in the continuum calculations are scaled with temperature.

## Introduction

In order to fully understand the thermodynamics of protein folding and protein−protein interactions, it is important to know the basic thermodynamic characteristics of the various forces, such as charge−charge and hydrophobic interactions, that drive these processes. A comprehensive thermodynamic characterization of these fundamental types of interactions requires, in turn, not only knowledge of the attendant change in free energy ($\Delta G$) but also the changes in the enthalpy ($\Delta H$), the entropy ($\Delta S$), and the heat capacity ($\Delta C_p$) throughout the temperature range of interest. Since there are "extremophilic" organisms that can survive and thrive at temperatures of 0 °C[1−3] and others that live happily at ∼100 °C,[4−7] a comprehensive understanding of biomolecular thermodynamics in aqueous solution requires us to consider a temperature range of 0° to 100 °C.

Experimental methods such as differential scanning calorimetry and isothermal titration calorimetry have been widely used to study the thermodynamics of protein folding[8−14] and protein−protein interactions,[15−21] respectively, and correlations of the resulting data with structural characteristics of proteins have allowed very useful estimates

* Corresponding author e-mail: adrian-elcock@uiowa.edu.

10.1021/ct1000704 © 2010 American Chemical Society
Published on Web 03/01/2010

to be obtained of, for example, the heat capacity change associated with the burial of hydrophobic surface area.[8,22−26] Given, however, that many different types of interactions act simultaneously to determine the thermodynamic properties of proteins, it can be difficult to unambiguously resolve the exact contribution made by a particular type of interaction.[25] One profitable way to circumvent this problem is to apply the same kinds of experimental methods to the study of small molecules that are representative of the components of proteins.[8,27−38] Even here, however, there can be difficulties of interpretation; it is not completely clear, for example, how good a model the commonly used N-methylacetamide is as a model of the polypeptide backbone of proteins.[38]

An alternative but complementary approach to direct experimentation is to use molecular simulation methods to directly measure the thermodynamic features of chosen types of interactions. Although computational results are always subject to concerns about the quality of the underlying force fields,[39,40] the simulation approach has the significant advantage of allowing contributions made by specific types of interactions to be resolved in a highly detailed and structurally unambiguous way: thermodynamic profiles can, for example, be obtained as a function of the distance between the interacting groups in a way that would be difficult, if not impossible, to achieve experimentally. A very large number of molecular simulation studies have already addressed the thermodynamics of hydrophobic associations in this way,[41−56] an abundance of literature that reflects the hydrophobic interaction's apparently dominant role in driving protein folding.[57] The first computed free energy profile for the association of two hydrophobic molecules in explicit solvent was reported many years ago.[58] Subsequent studies[51,55] explored the temperature dependence of the interaction and found, in common with experiment,[59,60] that it became stronger as the temperature increased. Only comparatively recently however have simulation studies been conducted at a sufficient number of different temperatures to allow the $\Delta C_p$ of hydrophobic association to be computed precisely.[50,52−54] Continuing this trend, a comprehensive study of the force field dependence of the $\Delta C_p$ for hydration and association of hydrophobic atoms has recently been reported,[55] showing that, while the various simulation models lead to quite different magnitudes of the computed $\Delta C_p$, they all correctly reproduce the signs of these $\Delta C_p$'s. That a qualitatively correct reproduction of the sign of the $\Delta C_p$ for hydrophobic association is a robust prediction of explicit-solvent simulation studies is quite important, since the *negative* $\Delta C_p$ is perhaps *the* defining, thermodynamic characteristic feature of protein folding.[61]

In contrast to the wealth of simulation studies that have examined the thermodynamics of the hydrophobic interaction, far less attention has been paid to using explicit-solvent molecular simulations to obtain *complete* thermodynamic characterizations of the other types of interactions that are present in proteins. In particular, very few studies have addressed the thermodynamics of attractive charge−charge interactions beyond measuring the free energy at, for example, 25 °C.[62] Such interactions are, however, likely to be of special interest from a thermodynamic perspective given the apparently critical role of salt bridges (i.e., oppositely charged ion pairs) in the adaptation of proteins to stability at high temperatures.[4−7] Free energy profiles for the association of a $Na^+$:$Cl^-$ ion pair were first computed using molecular dynamics (MD) simulations many years ago, and a comparison of results obtained at 25° and 100 °C indicated that the higher temperature favored the formation of a contact ion pair.[62] More recently, MD simulations have been used to simulate the diffusive association of lysine and glutamate residues at 25°, 50°, 75°, and 100 °C and to compute free energy profiles for the formation of a prototypical salt bridge at the same temperatures;[63] again, these studies showed the salt bridge interaction to be stabilized by increasing temperature. All of these studies fall far short however of providing the coverage of different temperatures necessary to obtain the $\Delta C_p$ associated with the formation of salt bridge interactions. In fact, the only works that we are aware of that have applied molecular simulation methods to explore heat-capacity-related properties of charged groups in aqueous solution are the pioneering studies conducted by the Sharp group aimed at understanding the origins of the $\Delta C_p$ of *hydration*.[64−68] These authors used explicit-solvent molecular dynamics (MD) simulations to identify structural features of the hydration shells of hydrophobic and charged atoms that correlate with their opposite signs of the $\Delta C_p$ of hydration observed experimentally; since all of their MD simulations were performed at a single temperature, however, they did not explicitly compute the $\Delta C_p$ directly from MD simulations.

This work describes the use of MD simulations to obtain a complete thermodynamic characterization of the association of a model salt bridge, the acetate−methylammonium pair, in explicit water and compares its thermodynamic features with those of a model hydrophobic association, the methane−methane interaction, for which we also report new results. Given the demonstration that the magnitudes of $\Delta C_p$ estimates can vary significantly depending on the water model used,[55] we have performed complete sets of simulations of both the acetate−methylammonium and methane−methane interactions with two popular water models, TIP3P[69] and TIP5P;[70] these models produced, respectively, the smallest and largest estimates of the $\Delta C_p$ of the hydrophobic interaction in the recent wide-ranging study referred to above.[55] The results reported here indicate that both water models reproduce the key qualitative feature that the $\Delta C_p$ for the formation of salt bridge and hydrophobic interactions are positive and negative, respectively, although, as expected, they differ significantly in their magnitudes. In addition, the results demonstrate that MD simulations in which associating molecules are allowed to freely diffuse can produce data of sufficient precision to allow $\Delta C_p$ to be reliably computed and provide a set of "gold standard" explicit-solvent data against which to compare the predictions of a commonly used implicit-solvent model based on continuum electrostatics.

## Methods

**Simulation Setup.** Molecular dynamics (MD) simulations were conducted at 11 independent temperatures in the range from −12.5 to 112.5 °C: −12.5°, 0°, 12.5°, 25°, 37.5°, 50°, 62.5°, 75°, 87.5°, 100°, and 112.5 °C using the GROMACS v3.3 software.[71,72] All simulations contained either one acetate molecule and one methylammonium molecule or two methane molecules immersed in a 25 × 25 × 25 Å box of water molecules; separate simulations were performed using the TIP3P[69] and TIP5P[70] water models. As in our previous work,[73] the partial charges and van der Waals parameters for the acetate and methylammonium were adapted from those assigned to glutamate and lysine respectively in the OPLS-AA parameter set.[74] A 10 Å cutoff was used for van der Waals and short-range electrostatic interactions, and the Particle Mesh Ewald (PME) method was used to describe all long-range electrostatic interactions.[75] Covalent bonds were constrained with the LINCS algorithm,[76] enabling a 2 fs time step to be used. Simulations were performed in the NPT ensemble, with the pressure being maintained at 1 atm with the Parrinello−Rahman barostat[77] and the temperature being maintained at the desired value with the Nosé-Hoover thermostat.[78,79] Prior to MD, energy minimization was carried out for 100 steps using the steepest-descent algorithm. Each system was then equilibrated for 10 ns before a production simulation lasting 500 ns was conducted, during which atomic coordinates were saved at 0.1 ps intervals, producing 5 million structural snapshots per simulation for analysis. Since all simulations at all temperatures were run for 500 ns, it was possible to obtain converged free energy estimates in almost all cases; the only exceptions were for desolvation barrier regions with the TIP5P water model at <12.5 °C: this water model has a tendency to freeze at low temperatures,[80] which in turn drastically decreases the number of diffusive encounters of the two solutes sampled during each simulation period (Figure S1, Supporting Information).

**Free Energy Surface Construction.** Two different kinds of free energy surfaces (FESs) were used to describe the association thermodynamics of the molecule pairs. The simplest kind was a one-dimensional FES (also known as a "potential of mean force"; PMF) constructed from histograms of the intermolecular distances extracted from the simulation snapshots.[63] For the acetate−methylammonium pair, this distance was defined in terms of the carboxyl carbon of the acetate molecule and the amino nitrogen of the methylammonium molecule; for the methane−methane pair, the separation distance was defined as the distance between the two carbons. The second kind of FES, used exclusively for the acetate−methylammonium system, was a more detailed, two-dimensional (2D) free energy surface constructed from histograms of the charge−charge separation distance defined above, and the hydrophobic−hydrophobic separation distance, defined as the distance between the two methyl groups of the two molecules.[73] Each histogram ranged from 2.1 to 28 Å, with bins of 0.1 Å width. As in our previous work, the solute−solute configurational entropy term,[73] which always strongly favors the dissociated state, was removed from consideration by comparing the MD-derived 2D histogram with a reference 2D histogram constructed by 100 million random placements of the two molecules into the same simulation box. The MD-derived histogram (which describes the solute−solute distribution obtained when the two solutes interact with each other in water) and the reference histogram (which describes the distribution expected when the two solutes are noninteracting in a vacuum) are used together to calculate the (excess) free energy of interaction for the *i*th 1D FES bin according to the equation: $\Delta G°(i,j) = -\mathrm{RT}\ \ln[P_{\mathrm{interacting}}(i)/P_{\mathrm{noninteracting}}(i)]$ or the *i*,*j*th 2D FES bin according to the equation $\Delta G°(i,j) = -\mathrm{RT}\ \ln[P_{\mathrm{interacting}}(i,j)/P_{\mathrm{noninteracting}}(i,j)]$, where $P$ indicates the probability of finding the two solutes at the separation distance(s) covered by the histogram bin. The resulting relative 2D-FESs were placed on an absolute scale using a protocol similar to the one we previously described:[73] free energy offsets were obtained from a linear regression between the MD-derived and continuum electrostatic-derived direct interaction free energies performed using histogram bins in which both the intercharge and intermethyl distance were between 10 and 15 Å (full details of the continuum solvent calculations are provided below).

As shown in our previous work, one complication of the use of PME electrostatics in MD simulations is that long-range electrostatic interactions between the solutes occur not only "directly" (i.e., between the two copies that are closest to each other) but also "indirectly" (i.e., with periodic images), and if no account is taken of these long-range indirect interactions, an incorrect view of the long-range direct interaction can be obtained.[73] In our previous work, we corrected for the presence of the periodic, indirect interactions by use of Poisson calculations in which two additional "layers" of solute images were included. In the present work, we have used a simpler but better protocol that makes use of the continuum electrostatics program DelPhi:[81,82] this program has the ability to rapidly perform Poisson calculations both with and without periodic boundary conditions. In fact, DelPhi's speed was found to be sufficient enough that it could be used to perform calculations on all 5 million structural snapshots from each MD simulation. The indirect energy for each snapshot was therefore obtained by taking the difference of the electrostatic energy computed from two near-identical Delphi calculations: one of the system with periodic boundary conditions applied in all three dimensions and one without periodic boundary conditions; importantly, all other parameters, including the grid mapping of the molecules, were identical in both calculations. The total number of solutions of the Poisson equation obtained during this analysis therefore amounted to 220 million. In all calculations, the grid spacing was set to 1 Å, and the molecule dielectric constant was set to 2; the solvent dielectric constant was set to the appropriate value for water at the temperature of interest (see below).[83] It is to be noted that the comparatively coarse grid spacing of 1 Å is acceptable for the calculations described above owing to the fact that their purpose is to calculate comparatively long-range electrostatic interactions between periodic images: although the use of a coarse grid would certainly lead to errors in calculations of the *direct* interaction between the two molecules, these errors would exactly cancel when calculations performed in nonperiodic and periodic conditions are compared. Having calculated indirect energy contributions for all 5 million snapshots obtained at each temperature, 2D-FESs describing these contributions could be constructed and subtracted from the 2D-FESs computed directly from the MD simulation data: the final, resulting 2D-FESs shown in the Results therefore describe only the thermodynamics of direct interaction between a single pair of interacting molecules.

**Thermodynamic Characterization.** To obtain the additional thermodynamic parameters $\Delta H$, $\Delta S$, and $\Delta C_p$, the $\Delta G$ values obtained at the different temperatures were fit to the Gibbs−Helmholtz equation:[50,52]

$$\Delta G(T) = \Delta H_0 + \Delta C_P(T - T_0) - T\Delta S_0 - T\Delta C_P \ln(T/T_0)$$

where $T_0$ is a reference temperature (chosen in this case to be 50 °C since it is at the center of the temperature range studied) and $\Delta H_0$ and $\Delta S_0$ are the respective interaction enthalpies and entropies at that temperature. Nonlinear fits of the simulation data to the above equation were conducted independently for all histogram bins by minimizing the mean square deviation between the $\Delta G$ values obtained from the MD simulations and the $\Delta G$ values predicted by the Gibbs−Helmholtz equation, allowing $\Delta C_p$, $\Delta H_0$, and $\Delta S_0$ to be free parameters. In the case of the acetate−methylammonium system, since these fits were performed independently for each bin in the 2D-FES, it was possible in turn to construct two-dimensional surfaces for $\Delta C_p$, $\Delta H_0$, and $\Delta S_0$ (see the Results). All fits were conducted with Microsoft Excel (http://office.microsoft.com). As written above, and as used in previous simulation studies,[50,52,55] $\Delta C_p$ is assumed to be independent of temperature. In the present study, however, we found that the precision of the simulation data, when plotted as a function of a single distance, was sufficiently high that it was also possible to discern a meaningful temperature dependence to $\Delta C_p$ by assuming a simple linear dependence on temperature, i.e., $\Delta C_p(T) = \Delta C_{p,0} + \alpha T$ (see the Results).

**Continuum Solvation Calculations.** To test whether the acetate−methylammonium 2D-FESs obtained from costly, explicit-solvent MD simulations could be reproduced by a faster implicit-solvent modeling scheme, additional calculations were performed with a combined continuum electrostatic + surface-area-based hydration model that one of us previously used[84,85] at the temperatures −12.5°, 25°, 62.5°, and 100 °C. "Direct" electrostatic contributions to association were computed using the Poisson−Boltzmann electrostatics program UHBD[86] following a protocol very similar to that used in our previous work.[73] Briefly, 15 structural snapshots from each bin on the 25 °C 2D-FES (obtained with the TIP3P water model) were randomly selected, and each was used to perform a series of Poisson calculations of (a) the methylammonium alone, (b) the acetate alone, and (c) the two molecules together, in the aqueous phase at each temperature of interest. All Poisson calculations were conducted in two stages: first, a 50 × 50 × 50 grid of spacing 1 Å was used, with boundary potentials assigned using the Coulombic approximation, and then a second, "focusing" calculation using a 100 × 100 × 100 grid of spacing 0.25 Å was performed. In all cases, the dielectric constants of the solutes were set to 2, while that of the solvent was set to the appropriate value for water at the temperature of interest, e.g., 78.45 for 25° and 55.57 for 100 °C.[83] The atomic charges used in all calculations were the same as those used in the MD simulations, and the atom radii were obtained from a previous extension to the PARSE parameter set[87] made by one of us to allow the parameter set's use at a wide range of temperatures.[84] In the extended PARSE model, the atomic radii are temperature dependent, increasing in size through the use of a radius scaling factor (RSF) as the temperature increases: different RSFs are applied to the atoms of the amino, carboxyl and methyl groups. As in the original PARSE scheme,[87] an additional surface-area based term was used to model nonelectrostatic contributions to association. In order to calculate these contributions at each temperature of interest, UHBD was first used to compute changes in solvent accessible surface areas due to association (with a probe radius of 1.4 Å), and these were then scaled by appropriate proportionality constants, $\gamma_{\text{aliphatic}}$ and $\gamma_{\text{polar}}$, parametrized in our previous work.[84] The total implicit-solvent interaction free energy at each bin of the 2D-FES was then obtained as the sum of these electrostatic and nonelectrostatic contributions: $\Delta G = \Delta G_{\text{elec}} + \Delta G_{\text{nonelec}}$.

## Results

**Temperature Dependence of the Free Energies of Association.** Unconstrained molecular dynamics (MD) simulations of acetate−methylammonium and methane−methane pairs in explicit solvent were performed at 11 temperatures with two different but commonly used water models. The resulting computed excess free energies of interaction, $\Delta G$, plotted versus the intermolecular distance, are shown for both types of association and for both types of water model in Figure 1 (note that different scales have been used for the two interaction types). As noted previously from simulations performed at 25 °C,[63] the computed $\Delta G$ for forming the salt bridge contact (left-hand panels) is significantly more favorable than for forming the methane−methane contact (right-hand panels); this remains true for all temperatures and with both water models. For both types of interaction, there is a general tendency for the $\Delta G$'s to become progressively more favorable as the temperature increases. This trend, while operative at all intermolecular distances, is especially apparent for the direct contact free energy minima, which are located at separation distances of 3.3 Å and 3.8 Å for the salt bridge and methane−methane interactions, respectively; the temperature dependencies of these contact minima are highlighted as insets in Figure 1. The latter plots also lead to two more significant observations. First, and most obviously, the plots of $\Delta G$ versus $T$ are nonlinear and curve in opposite directions for the hydrophobic and charge−charge interactions. Second, the extents of curvature are much greater for the simulations performed with the TIP5P water model than for the TIP3P model. As is considered in more detail later, these aspects of the plots report directly on the different $\Delta C_p$'s of interaction.

Performing global fits of the free energy data to the Gibbs−Helmholtz equation provides smoothed estimates of the $\Delta G$ of interaction at all temperatures (see the Methods section); the quality of such fits, for both the charge−charge and hydrophobic interactions, can be assessed by examining

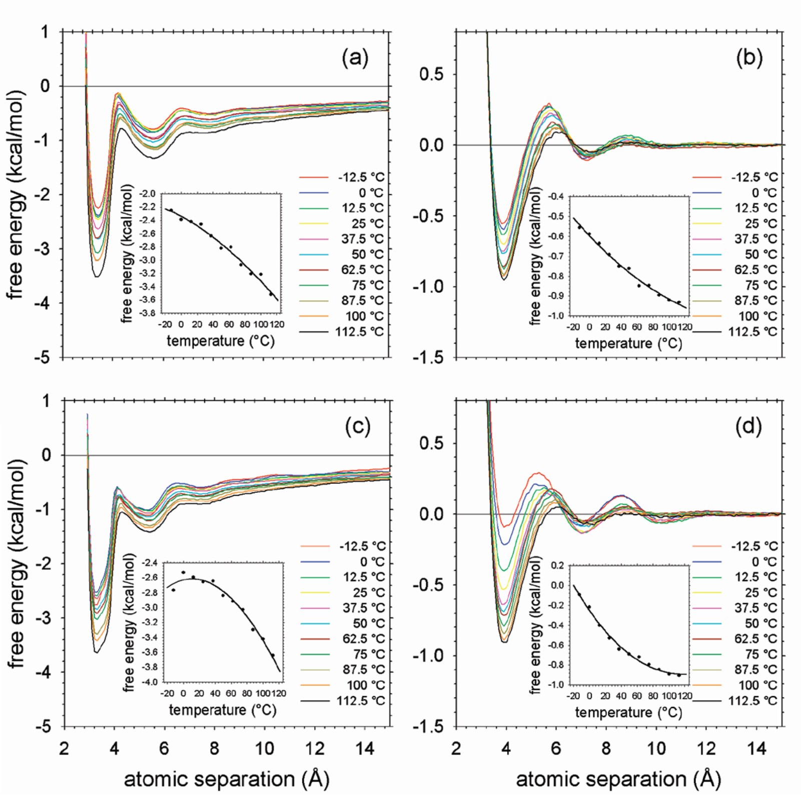


***Figure 1.*** Excess free energy of interaction of acetate−methylammonium (a, c) and methane−methane pairs (b, d) plotted as a function of the intermolecular distance for all temperatures in the range −12.5° to 112.5 °C. Panels a and b are for the TIP3P water model; panels c and d are for TIP5P. The insets show a close-up of the behavior of the global free energy (contact configuration) minima.

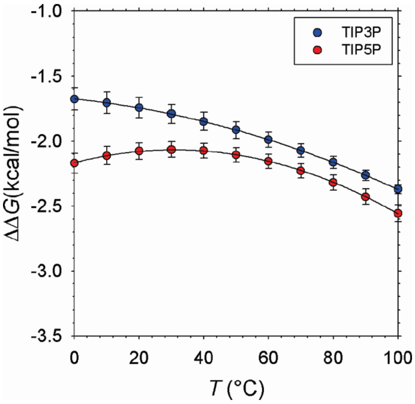


***Figure 2.*** Plot of the difference between the smoothed $\Delta G$ values for the direct charge−charge and hydrophobic contacts plotted versus temperature for TIP3P (blue circles) and TIP5P (red circles) water models. Error bars represent the standard deviation obtained from three independent Gibbs−Helmholtz fits: for the charge−charge contact, fits were performed using data at separation distances of 3.2, 3.3, and 3.4 Å; for the hydrophobic contact, fits were performed using data at separation distances at 3.7, 3.8, and 3.9 Å.

the solid lines of the insets to Figure 1. In Figure 2, we show the difference, $\Delta\Delta G$, between the smoothed $\Delta G$ values of the direct charge−charge and hydrophobic interactions as a function of the temperature for the two water models. The $\Delta\Delta G$ is negative at all temperatures due to the considerably greater affinity computed for the charge−charge interaction (see above). But a much more interesting result concerns the temperature dependence of this $\Delta\Delta G$. Between ~40° and 100 °C, both the TIP3P and TIP5P water models predict essentially identical behavior: the $\Delta\Delta G$ increases in magnitude as the temperature increases, indicating that, relative to the hydrophobic contact, the charge−charge contact becomes progressively stronger at higher temperatures.[63] Between 40° and 0 °C, however, there is a clear difference between the two water models: while the TIP3P curve continues to show a significant temperature dependence, the TIP5P curve flattens completely and predicts effectively no change in the *relative* strengths of charge−charge and hydrophobic interactions over this temperature range. As is considered in detail in the Discussion section, this result may have implications for understanding the relative roles played by hydrophobic and salt bridge interactions in proteins from psychrophilic, mesophilic, and hyperthermophilic organisms.

**Gibbs−Helmholtz Fits of the 1D Free Energy Functions.** Globally fitting $\Delta G$ data to the Gibbs−Helmholtz equation not only allows smoothed estimates of $\Delta G$ to be obtained but it also allows a deeper thermodynamic understanding to be obtained from examination of the fitted $\Delta H$, $\Delta S$, and $\Delta C_p$ values (see the Methods section). The resulting computed $\Delta H$'s of the charge−charge and hydrophobic interactions are plotted versus temperature in Figure 3a; corresponding plots of the $\Delta S$ are shown in Figure 3b. The different *qualitative* temperature dependences of the charge−charge and hydrophobic interactions are apparent from both figures: the slope of the $\Delta H$ for the charge−charge interaction

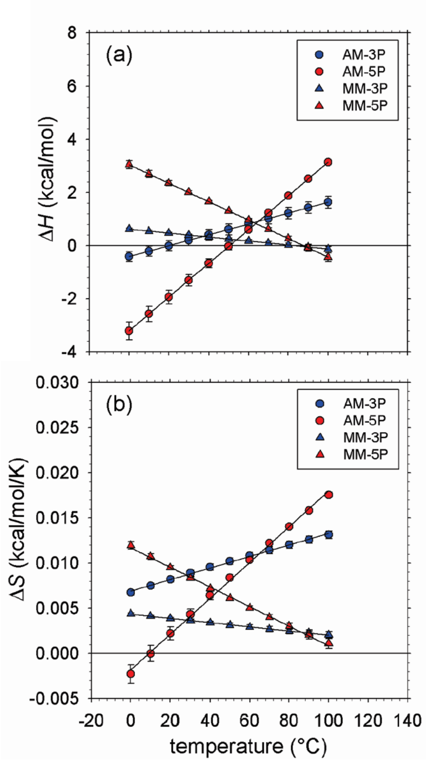


***Figure 3.*** Plots of $\Delta H$ (a) and $\Delta S$ (b) versus temperature for acetate−methylammonium (AM; circles) and methane−methane (MM; triangles) systems. Results for TIP3P (3P) and TIP5P (5P) water models are shown as blue and red symbols, respectively. Error bars were computed in the same way as described in the legend to Figure 2.

(circles), for example, is of opposite sign to the slope of the $\Delta H$ for the hydrophobic interaction (triangles). The significant *quantitative* differences between the results obtained with the two water models, on the other hand, can be seen by comparing the very different slopes of the red (TIP5P) and blue (TIP3P) symbols.

One issue that is especially notable in Figure 3a and b concerns the temperature at which the $\Delta H$ and $\Delta S$ of the *hydrophobic* interaction approach zero (triangles). Experimentally, it has been shown that the values of the hydration $\Delta H$, of unfolding of a wide range of proteins tend to zero at ∼90 °C, and that the corresponding $\Delta S$ values tend toward zero at 112−118 °C;[33,35,36,88−90] the origins of this effect, and what it tells us about the hydrophobic effect, have been the subject of considerable study.[33,35,36,91−95] Interestingly, the $\Delta H$ and $\Delta S$ values of the hydrophobic interaction computed with the TIP5P water model reach zero, or are extrapolated to reach zero, at ∼87 °C and ∼108 °C, respectively, both of which values are in very good agreement with the experiment. The $\Delta H$ and $\Delta S$ values obtained with the TIP3P water model, on the other hand, reach zero at ∼83 °C and ∼187 °C, which is qualitatively correct (in the sense that the zero-point temperature for $\Delta S$ is higher than that for $\Delta H$), but in poor quantitative agreement. On the basis of this comparison, therefore, one might conclude that the TIP5P water model gives a more realistic description of the effects of temperature on biomolecular association thermodynamics.

As noted above, the curvatures apparent in the insets of Figure 1 are indicative of a nonzero $\Delta C_p$ for the charge−charge and hydrophobic interactions. For the TIP5P water model, the $\Delta C_p$ values obtained from Gibbs−Helmholtz fits—in which $\Delta C_p$ is assumed to be temperature-independent—are +60 and −35 cal/mol/K for the charge−charge and hydrophobic contact interactions, respectively; for TIP3P, the corresponding numbers are +19 and −6 cal/mol/K, respectively. Both water models therefore correctly reproduce the fact that the $\Delta C_p$ for charge−charge and hydrophobic interactions, being dominated by hydration terms, will be of opposite sign,[24,26,30,96−98] with the TIP5P model, as expected, producing the larger estimates. For the hydrophobic interaction, the computed $\Delta C_p$ values can be compared with values derived by regression analyses of experimental protein folding thermodynamics data[8,22−24] and small-molecule solubility data.[99] The experimental data are typically expressed in a form normalized by the degree of buried nonpolar surface area; typical regressed values of the nonpolar $\Delta C_p$ contributions are −1.9,[8] −1.4,[22] −1.2,[23] and −2.1[24] J/mol/K/Å$^2$. Normalizing our computed estimates by the amount of surface area buried in the direct contact configuration (64.3 Å$^2$), we obtain $\Delta C_p$ contributions of −2.3 and −0.4 J/K/mol/Å$^2$ for TIP5P and TIP3P, respectively; as might have been anticipated, therefore, the two water models give estimates that straddle the corresponding experimental estimates.

Analysis of the computed $\Delta C_p$ behavior can in fact be taken a stage further by exploring potential temperature dependences of the charge−charge and hydrophobic $\Delta C_p$ values (see the Methods section); this has been done by repeating the Gibbs−Helmholtz fits under the assumption that $\Delta C_p$ is a *linear* function of temperature. For the TIP3P water model, the much smaller absolute values of the $\Delta C_p$'s make it difficult to be certain of any trend, but for TIP5P a clear temperature dependence is apparent: for all interaction distances less than 8 Å, the $\Delta C_p$ of the charge−charge interaction becomes progressively more positive as temperature increases while $\Delta C_p$ of the hydrophobic interaction becomes progressively more negative (Figure 4a). The temperature dependence of $\Delta C_p$ for both types of interaction at their contact distances are illustrated in Figure 4b and compared with Privalov and Makhatadze's experimental estimates[34] of the polar and nonpolar $\Delta C_p$ contributions made to protein folding in Figure 4c. The behavior obtained with the TIP5P water model is in surprisingly good qualitative agreement with that seen experimentally.

**A More Detailed View of the Acetate−Methylammonium Interaction.** A more detailed view of the effects of temperature on the interaction thermodynamics of the acetate−methylammonium pair offers an opportunity to examine the *simultaneous* operation of charge−charge and hydrophobic interactions.[73] This can be done by constructing two-dimensional free energy surfaces (2D-FESs) in which the *x* coordinate is the charge−charge distance (the $C_{carboxyl}$−$N_{amino}$ distance) and the *y* coordinate is the distance between the hydrophobic groups (the $C_{methyl}$−$C_{methyl}$ distance). Representative two-dimensional free energy surfaces (2D-FESs) obtained with the TIP5P water model are shown

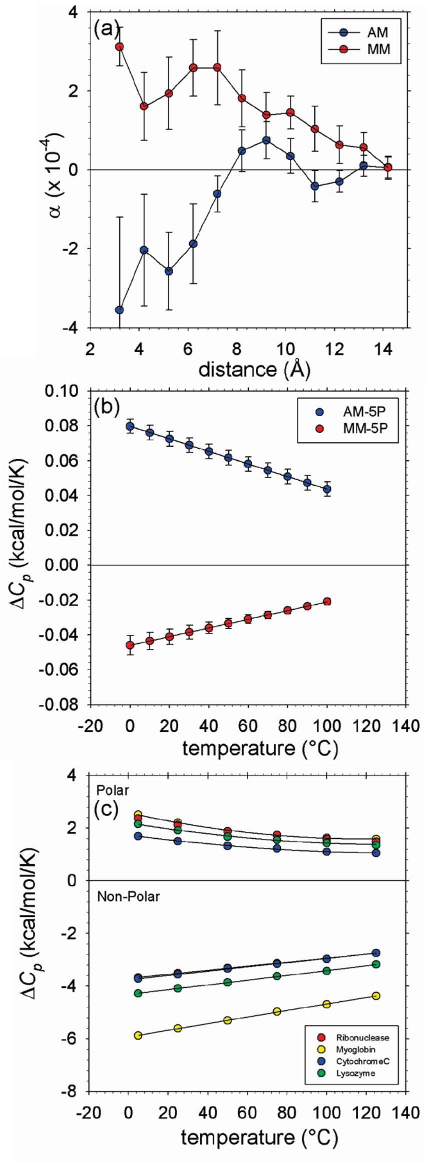


***Figure 4.*** (a) Temperature dependence of $\Delta C_p$, α, plotted versus intermolecular distance for acetate–methylammonium (AM; blue circles) and methane–methane (MM; red circles) systems, data obtained with the TIP5P water model. (b) Computed $\Delta C_p$ for the direct charge–charge (AM; blue circles) and hydrophobic (MM; red circles) contacts versus temperature, data obtained with the TIP5P water model. (c) Experimental $\Delta C_p$ versus temperature for cytochromeC (blue circles), ribonuclease (red circles), lysozyme (green circles), and myoglobin (yellow circles) folding, data taken from ref 34. Error bars shown in a represent the standard error obtained from the Gibbs–Helmholtz fit performed in SigmaPlot.[122] Error bars for b were computed in the same way as described in the legend to Figure 2.

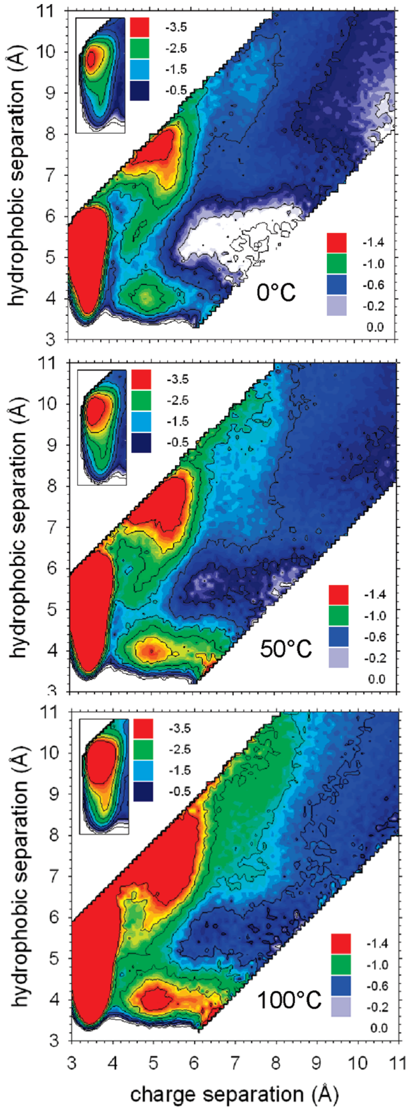


***Figure 5.*** 2-Dimensional (excess) interaction free energy surfaces (2D-FESs) for acetate–methylammonium for three temperatures, data obtained with the TIP5P water model. The insets show close-ups of the 2D-FES in the region of the charge–charge contact, replotted on energy scales that allow identification of the global minimum.

for 0°, 50°, and 100 °C in Figure 5; complete results for all 11 temperatures and for both water models are provided in Figures S2 and S3 (Supporting Information). On each 2D-FES, there are three pronounced minima. The first, the global minimum configuration, is the direct charge–charge contact (i.e., a conventional salt bridge) located at coordinates (*x*, 3 Å; *y*, 5 Å) on the 2D-FES; the two other (local) minima are a solvent-separated charge–charge interaction at coordinates (*x*, 5.5 Å; *y*, 7 Å) and a broad methyl–methyl hydrophobic contact at (*x*, 5–6 Å; *y*, 4 Å). The same features appear on the 2D-FESs obtained with the TIP3P water model, with the only notable differences—such as a broader minimum for the methyl–methyl contact—being caused by TIP5P having a lower free energy barrier to the dissociation of the charge–charge contact (apparent from comparing Figure 1a and c; see also Figure S4, Supporting Information).

As anticipated from the behavior of the one-dimensional free energy functions shown in Figure 1, increasing the temperature stabilizes all three of the local minima on the 2D-FES. The differing degrees of stabilization of the three local minima, however, become more apparent in the form of 2D-FES-difference maps obtained by subtracting 2D-FESs at 100 and 25 °C (Figure 6a,b). From such plots, it is apparent that for *both* water models the direct charge–charge interaction is much more strongly stabilized by increasing temperature than the other modes of interaction. Interestingly, a plot of the data points from the TIP3P 2D-FES-difference

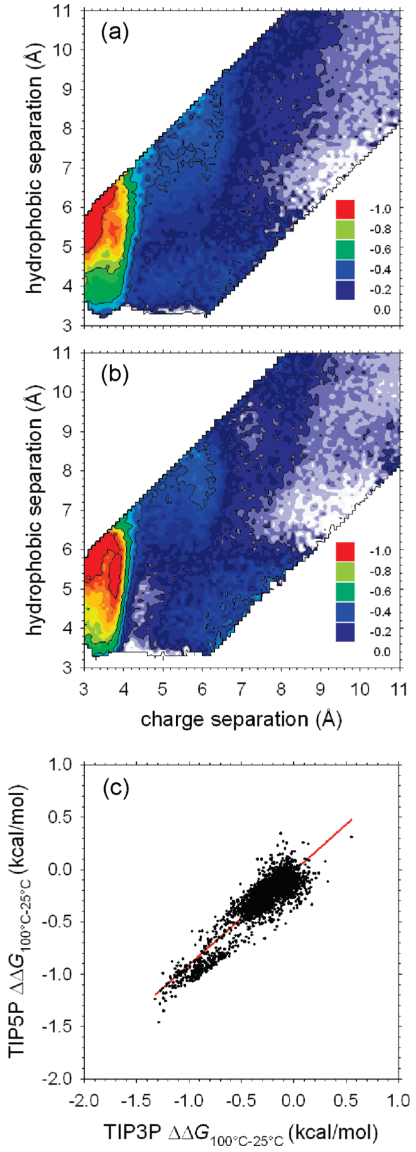


***Figure 6.*** (a) 2D-FES difference surface ($\Delta\Delta G_{100-25°C}$) for the TIP3P water model. (b) 2D-FES difference surface ($\Delta\Delta G_{100-25°C}$) for TIP5P. (c) Correlation of data points from the TIP5P 2D-FES difference surface with those from the corresponding TIP3P 2D-FES difference surface; only data points with a charge−charge separation between 2.9 and 20 Å are plotted. The red line is the linear regression line with slope = 0.8934, intercept = −0.0182, and $R^2$ = 0.8306.

map versus the corresponding data points from the TIP5P 2D-FES-difference map shows—for the temperature range 25° to 100 °C—a high degree of correlation ($R^2$ = 0.83), with a slope (0.89) indicating a slightly greater temperature dependence for the TIP3P model (Figure 6c). As explored in detail below, these 2D-FES-difference maps provide a critical test for continuum solvation models intended to describe the temperature dependence of biomolecular interactions at high temperatures.

As with the 1D free energy functions, it is possible to extract further thermodynamic information by globally fitting the 2D-FESs to the Gibbs−Helmholtz equation: doing so allows 2D surfaces for the $\Delta H$, $\Delta S$, and, most importantly, the $\Delta C_p$ of the interaction to be derived (see the Methods section). Examples of the kinds of fits that are obtained are shown in Figure S5 (Supporting Information) for the three major free energy minima on the TIP5P 2D-FES; similar results obtained with the TIP3P model are shown in Figure S6 (Supporting Information). As might be expected, the best fits to the Gibbs−Helmholtz equation are obtained in those regions on the 2D-FESs that are most frequently sampled during the MD simulations: for example, in the case of the three regions shown in Figures S5 and S6 (Supporting Information), the fit is clearly better for the global minimum configuration. This connection between sampling efficiency and quality of the Gibbs−Helmholtz fits is shown in a 2D surface representation for both water models in Figure S7 (Supporting Information): from these plots, the tendency for errors to be greatest in the regions of the surface that are poorly sampled is clear, especially in those regions corresponding to desolvation barriers. In fact, for certain parts of the 2D-FES, the adequacy of sampling was questionable for the TIP5P water model at the three lowest temperatures studied (−12.5°, 0°, and 12.5 °C); these 2D-FESs were therefore omitted from the global Gibbs−Helmholtz fits. Despite this caution—which, it should be noted, does *not* apply to the free energies when plotted as a function of a *single* dimension (Figure 1)—the $\Delta G$ values computed directly from these low temperature simulations were usually in good agreement with the $\Delta G$ values extrapolated from the Gibbs−Helmholtz fits to the 8 higher temperatures (see open symbols in Figure S5).

Two-dimensional surfaces illustrating the $\Delta H$ and $T\Delta S$ of interaction at a number of temperatures are shown in Figures S8 and S9 (Supporting Information); the more interesting quantity to examine however is the heat capacity, $\Delta C_p$, which is shown in a 2D representation for both water models in Figure 7. The TIP3P water model (Figure 7a) produces a very slightly positive $\Delta C_p$ of interaction over much of the 2D surface. The TIP5P water model (Figure 7b), in contrast, produces a strongly positive $\Delta C_p$ for the direct charge−charge contact, and a strongly negative $\Delta C_p$ for the hydrophobic contact (which matches well with the behavior seen in the methane−methane simulations). With the TIP5P model, therefore, it is possible to discern distinct modes of interaction between the same two molecules that have qualitatively different $\Delta C_p$ behaviors; moreover, effecting such a qualitative change in the $\Delta C_p$ requires only a shift in relative orientation of the two molecules of a few Ångstroms.

**Comparison with Implicit Solvation Calculations.** The availability of explicit solvent free energy surfaces computed over a range of temperatures provides an excellent opportunity to test the ability of implicit (continuum) solvent models to capture temperature dependent effects. In what follows, therefore, the interaction thermodynamics obtained from MD are compared with corresponding results obtained using the current "gold standard" implicit solvent model, the Poisson(−Boltzmann) method.[100−102] Two sets of Poisson calculations were carried out: one in which the only temperature-dependent parameter in the calculations was the solvent dielectric constant and one in which, additionally, the atomic radii were adjusted by a temperature-dependent radius scaling factor (RSF) empirically derived previously[84] to reproduce the experimental hydration free energies of

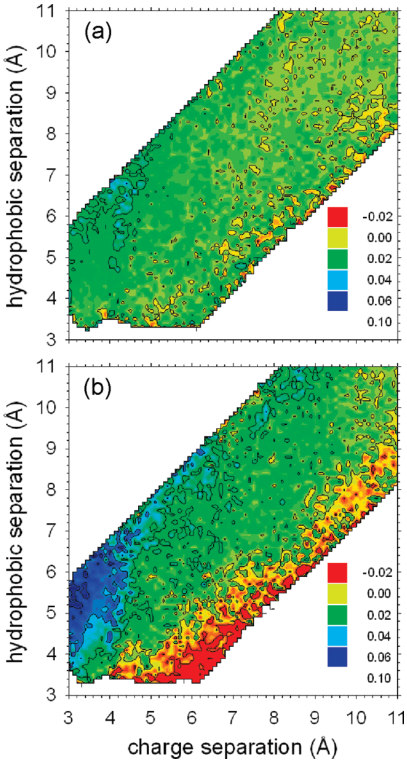


***Figure 7.*** 2D surfaces showing the heat capacity change of interaction, $\Delta C_p$, for the (a) TIP3P and (b) TIP5P water models.

amino acids over a wide range of temperatures. Nonpolar contributions to the acetate−methylammonium interaction were calculated using a temperature-dependent solvent accessible surface area (SASA) term (see the Methods section).

2D-FES-difference maps showing the effects of changing the temperature from 25° to 100 °C are shown for both Poisson calculation protocols in Figure 8a and b; these implicit-solvent surfaces should be compared with the explicit-solvent surfaces shown in Figure 6a and b. Such a comparison indicates that treating the atomic radii as temperature dependent with the RSF results in much better agreement with the explicit solvent MD results, especially in terms of describing the temperature dependence of the direct charge−charge interaction. In fact, linear regression of the Poisson-computed free energy differences with those obtained from the TIP5P simulations shows significant improvements in the slope, the intercept, and the $r^2$ value when the RSF is included (see Figure 9a and b, respectively). A qualitatively identical finding is obtained when the implicit solvent results are compared instead to the TIP3P results (Figure S10, Supporting Information).

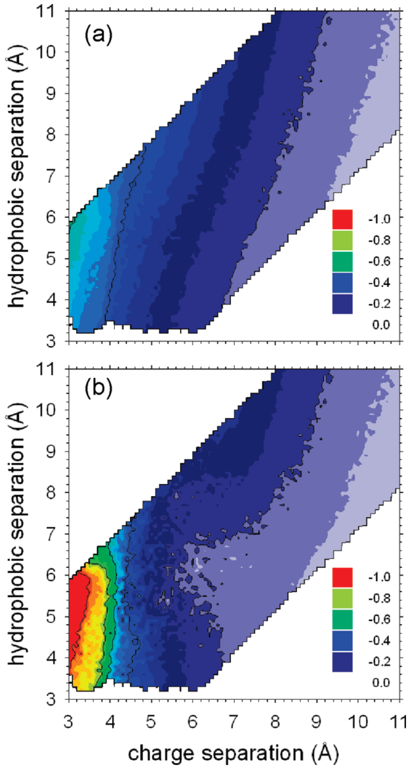


***Figure 8.*** 2D-FES difference surfaces ($\Delta\Delta G_{100-25°C}$) obtained from Poisson−Boltzmann calculations performed (a) without radius scaling factor (RSF) and (b) with RSF.

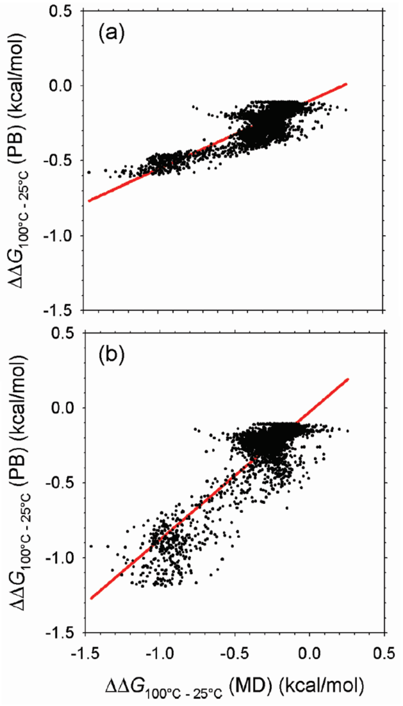


***Figure 9.*** Correlation of $\Delta\Delta G_{100-25°C}$ calculated from Poisson−Boltzmann calculations and from TIP5P MD simulations. (a) PB calculations performed without RSF: slope = 0.4513, intercept = −0.1059, $r^2$ = 0.7313. (b) PB calculations performed with RSF: slope = 0.8506, intercept = −0.0263, $r^2$ = 0.7873.

## Discussion

As noted in the Introduction, the studies reported here have been conducted with two water models that have been chosen to provide what are thought to be extreme descriptions of the likely effects of temperature on biomolecular interactions.[55] While significant differences are certainly found between the two models (see below), it is striking that they make very similar predictions of the temperature's effects on the free energies of the hydrophobic and charge−charge interactions at medium-to-high temperatures. Increasing the temperature from 25° to 100 °C changes the $\Delta G$ of the direct charge−charge interaction by −0.76 and −0.77 kcal/mol with the TIP3P and TIP5P models, respectively (Figure 1a,c) and changes the $\Delta G$ of the methane−methane interaction by −0.23 and −0.36 kcal/mol with TIP3P and TIP5P, respectively (Figure 1b,d). Importantly, as we discuss next, this close correspondence between the two explicit solvent water models at medium-to-high temperatures allows unambiguous conclusions to be drawn about the use of implicit solvent (Poisson) calculations for investigating the effects of high temperatures on biomolecular interaction thermodynamics.

Poisson and Poisson−Boltzmann calculations are widely used to provide insights into electrostatic contributions to biomolecular interactions[100−102] and have been especially exploited to investigate the contributions of salt bridges to the stability of proteins from hyperthermophilic organisms.[85,103−108] A number of the latter studies have attempted to compare salt bridge interaction thermodynamics at low and high temperatures by comparing calculation results[105,107,108] obtained with the solvent dielectric constant set to the appropriate experimental value for water (the dielectric constant of water changes from 87.9 at 0 °C to 55.6 at 100 °C[83]). Ideally, altering the solvent dielectric constant in this way would be the *only* change required in order for implicit solvent models to capture accurately the effects of changing temperature on interaction thermodynamics. Previous work carried out by one of us,[84] however, found that the temperature dependence of amino acid *hydration* free energies was systematically underestimated (compared with the experiment) when this approach was followed; in order to obtain agreement with the experiment, it was found necessary to adjust the atomic radii by an empirically determined radius scaling factor[84] (RSF). Since changes in hydration play such a critical role in determining the thermodynamics of interactions between charged residues,[85,103] it is to be anticipated that the requirement for adjustable atomic radii might reappear in attempts to match temperature dependent changes in *association* free energies. The explicit solvent MD results reported here have provided the opportunity to examine this issue. While the overall agreement between the 2D-FES-difference maps obtained from explicit solvent (Figure 6) and implicit solvent (Figure 8) calculations is far from perfect, the overall trend is pretty clear: the free energy change obtained when the solvent dielectric constant is the *only* parameter changed in the implicit solvent calculations is much smaller than that obtained from explicit solvent simulations, but the free energy change obtained when the RSF is used is much closer in magnitude. With regard to continuum solvation calculations, therefore, the basic conclusion to be drawn from the present study is the following. On the basis of two independent lines of evidence—comparisons with (a) the experimental temperature dependence of amino acid hydration free energies[84] and (b) the MD-simulated temperature dependence of salt bridge association free energies (shown here)—a strong case can be made that atomic radii should be adjusted with a RSF in Poisson or Poisson−Boltzmann calculations aimed at modeling temperature dependent changes in biomolecular thermodynamics.

As noted above, this conclusion can be drawn with confidence owing to the fact that the quite different TIP3P and TIP5P water models make essentially identical predictions about the overall magnitude of $\Delta G$ changes in the range 25° to 100 °C; another way of saying this is that the computed first derivative of $\Delta G$ with respect to temperature (in this temperature range) is very similar for both models. Where the two water models produce quite different results is in the more subtle quantity $\Delta C_p$, which describes the second derivative of $\Delta G$ with respect to temperature (i.e., its curvature). The next question to ask is of course which of the $\Delta C_p$ predictions, if any, should be believed? Owing to the fact that the TIP5P water model was specifically devised in order to correct inadequacies in the treatment of temperature effects on water's density,[70] one might immediately anticipate that it would also provide the more accurate description of temperature effects on interaction thermodynamics. Certainly, we know from the work of others that TIP5P provides a good description of the temperature dependence of water's dielectric constant. Specifically, it has been shown to produce dielectric constant values of 82 and 60 at 25 and 100 °C, respectively,[70] which compare well with the experimental values of 78.5 and 55.6;[83] unfortunately, we cannot make a corresponding comparison for TIP3P as we have been unable to find estimates of its dielectric constant at 100 °C. An additional piece of quantitative evidence in favor of TIP5P is reported here: we find that the temperatures at which $\Delta H$ and $\Delta S$ of the hydrophobic interaction equal zero are in good accord with experimental estimates for the TIP5P model but are not for TIP3P. Other observations provide suggestions, but not outright proof, that TIP5P's description is better than that of TIP3P. With the acetate−methylammonium system, for example, only the TIP5P model provides clear evidence of differently signed $\Delta C_p$'s for the charge−charge and methyl−methyl interactions. With both systems, only the TIP5P model gives a discernible trend in the temperature dependence of the $\Delta C_p$'s of interaction, and this, in turn, is in good qualitative correspondence with the trend seen in the experimental data.[34]

That said, there are other aspects of behavior of the two water models that lead to more equivocal conclusions. For example, a recent comprehensive simulation study of the heat capacity change, $\Delta C_p$, accompanying the *hydration* of methane (modeled with the OPLS united atom model) produced estimates of 145 and 265 J/K/mol for the TIP3P and TIP5P water models, respectively.[55] The experimental estimates of the same quantity range from 209 to 242 J/K/mol[99] and so lie somewhere in between the predictions of the two models. Similarly, as outlined in the Results section, comparison of the computed $\Delta C_p$ values of the hydrophobic interaction obtained with the two water models (−2.3 and −0.4 J/K/mol/Å$^2$ for TIP5P and TIP3P, respectively) shows that they lie on either side of estimates obtained from regressions of experimental data.[8,22−24] There are even some respects in which TIP5P is clearly not as good as TIP3P. For example, the freezing and boiling points reported recently for the TIP3P water model (269 and 357 K, respectively[109]) are considerably better than those reported in the same work for the TIP5P model (266 and 337 K, respectively;[109] the latter has also been independently estimated at 348 K[110]). In addition, the computed heat capacity of *pure* TIP5P water (29.2 J/K/mol) is in somewhat worse agreement with the experiment (18.0 J/K/mol) than are the heat capacities of the simpler TIP3P, SPC, and TIP4P water models ($C_p$ values of 20.0, 20.2, and 20.4 J/K/mol, respectively).[111] This latter result indicates, as noted by Mahoney and Jorgensen, that a better treatment of structural properties does not always lead to a better modeling of thermodynamic quantities.[70] It is certainly possible, therefore, that other water models might reproduce aspects of temperature-dependent thermodynamics somewhat better than TIP5P.

Before leaving the subject of $\Delta C_p$, it is worth noting that the same regression studies of experimental data referred to above have also indicated that, when expressed in units of buried surface area, the magnitude of the (positive) contribution to $\Delta C_p$ made by the burial of polar groups is considerably *smaller* than that of the (negative) contribution to $\Delta C_p$ made by the burial of nonpolar groups: the reported respective $\Delta C_p$ values for burial of polar vs nonpolar groups are +1.1 vs −1.9 J/mol/K/Å,[2,8] +0.7 vs −1.4 J/mol/K/Å,[2,22] +0.4 vs −1.2 J/mol/K/Å,[2,23] and +0.9 vs −2.1 J/mol/K/Å.[2,24] It is noticeable that the magnitude of the MD-computed $\Delta C_p$ for the charge−charge interaction is some 2−3 times *larger* than that of the hydrophobic interaction: with the TIP5P model, the $\Delta C_p$ for formation of the direct charge−charge contact is +60 cal/mol/K while that for formation of the hydrophobic (methane−methane) contact is −35 cal/mol/K. For TIP3P, the $\Delta C_p$ for formation of the direct charge−charge contact is +19 cal/mol/K, while that for formation of the hydrophobic (methane−methane) contact is −6 cal/mol/K. The most likely explanation of this apparent discrepancy is simply that, in the regressions of the experimental data, the "polar" contribution encompasses neutral, hydrogen bonding groups (especially the peptide backbone) in addition to charged groups; formation of neutral hydrogen bonding interactions, which have not been studied here, are likely to make much smaller contributions to heat capacity changes.

The qualitative difference between the $\Delta C_p$ values associated with the formation of charge−charge and hydrophobic contacts has consequences for their relative stabilities at both high and low temperatures (Figure 2). The finding that a salt-bridge interaction becomes progressively more stable than a hydrophobic interaction at high temperatures (∼40−100 °C) is not especially new, since we have found much the same result previously using both implicit[85] and TIP3P-based explicit solvent simulations.[63] It is nevertheless notable, however, that the same basic result is obtained with the TIP5P water model, despite the large quantitative differences between the TIP5P and TIP3P $\Delta C_p$ values. This indicates that the basic conclusions drawn previously[63] are not dependent on the water model. The thermodynamic behavior obtained in both the present and previous simulations provides an attractive explanation—though not the only one imaginable[104−107]—for the unusual abundance of salt bridge interactions in proteins from hyperthermophilic organisms.[4,112]

The more novel result of the present study is that, with the TIP5P water model, the relative strengths of salt-bridge and hydrophobic interactions are largely unchanged between 0° and 40 °C. Just as the preference of salt bridges for high temperatures has apparent implications for understanding the adaptation mechanisms of hyperthermophilic organisms,[4] this new finding may have implications for understanding organisms adapted to life at very low temperatures (psychrophiles). If it is indeed true that the relative stabilities of electrostatic and hydrophobic interactions remain essentially unchanged as the temperature drops to 0 °C, then there should be no selective advantage to accumulating or losing salt bridges at low temperatures: the relative numbers of salt bridges and hydrophobic interactions should therefore be very similar in mesophiles and psychrophiles. Since a number of crystal structures of proteins from psychrophilic organisms have recently become available, the accuracy of the above prediction can be directly examined. In fact, in the majority of cases that have been reported so far, the numbers of salt bridges in psychrophilic enzymes are indeed very similar to[113−115] or somewhat lower than[116−118] those found in their mesophilic homologues. One interesting exception that we know of is citrate synthase: the psychrophilic (and hyperthermophilic) versions of this enzyme have increased numbers of salt bridges relative to their mesophilic cousin, although in the case of the psychrophile, only *intra*-subunit salt bridges are found to be increased.[107] Exceptions are perhaps to be anticipated: obtaining unambiguous views of the adaptation mechanisms operating in psychrophiles is likely to be more difficult than for hypthermophiles since the former face the challenge of simultaneously retaining not only stability but also activity in their chosen environmental conditions.[1−3]

## Summary

The complete thermodynamic characterization of two types of molecular interactions in the range 0° to 100 °C shows that there are areas in which typical simulation water models are likely to produce essentially identical behavior and areas in which they will differ markedly. Both water models can capture the qualitative result that the $\Delta C_p$ for formation of salt bridge interactions is positive—which is something that has not been shown before—while the $\Delta C_p$ for formation of hydrophobic interactions is negative; the models differ drastically however in their predictions of the magnitude of the $\Delta C_p$. Both water models predict that on raising the temperature from ∼40° to 100 °C salt-bridge interactions are significantly more stabilized than are hydrophobic interactions. But they differ drastically in their predictions of what happens when the temperature drops from 40° to 0 °C. As noted above, the similarity of the models' predictions at high temperatures enables us to draw some firm conclusions regarding protocols for implicit solvent calculations at high temperatures; it also argues that for molecular dynamics simulations aimed at investigating biomolecular behavior at high temperatures the choice of water model may not be especially important. But the very significant differences observed at low temperatures on the other hand—which on balance appear to favor the TIP5P model—suggest that a closer examination of behavior at low temperatures may be important for force field development. Finally, it should be noted that, while we have explicitly compared the thermodynamic characteristics of hydrophobic and favorable charge−charge interactions here, these are not the only types of interactions to play important roles in determining biomolecular stability. It will in particular be of interest to explore similar issues for the thermodynamics of hydrogen bonding[119] and cation−$\pi$[120,121] interactions.

**Acknowledgment.** The authors are grateful to Dr. Andy S. Thomas for thoughtful discussions. This work was supported by a NSF CAREER award (#0448029) to A.H.E.

**Supporting Information Available:** Ten supporting figures: Time course of minimum distance between molecules versus time at the two lowest temperatures; 2D-FESs at all temperatures for TIP3P and TIP5P models; difference 2D-FESs between TIP5P and TIP3P at two temperatures; Gibbs−Helmholtz fits at three minima for TIP3P and TIP5P models; 2D surface showing errors in Gibbs−Helmholtz fits; 2D surfaces showing $\Delta H$ and T$\Delta S$ at two temperatures for TIP3P and TIP5P models; correlation of PB and TIP3P free energy differences with and without RSF. This information is available free of charge via the Internet at http://pubs.acs.org/.

## References

(1) Cavicchioli, R.; Siddiqui, K. S.; Andrews, D.; Sowers, K. R. *Curr. Opin. Biotechnol.* **2002**, *13*, 253–261.

(2) Feller, G.; Gerday, C. *Nat. Rev. Microbiol.* **2003**, *1*, 200–208.

(3) Siddiqui, K. S.; Cavicchioli, R. *Annu. Rev. Biochem.* **2006**, *75*, 403–433.

(4) Sterner, R.; Liebl, W. *Crit. Rev. Biochem. Mol. Biol.* **2001**, *36*, 39–106.

(5) Vieille, C.; Zeikus, G. J. *Microbiol. Mol. Biol. Rev.* **2001**, *65*, 1–43.

(6) Kumar, S.; Nussinov, R. *Cell. Mol. Life Sci.* **2001**, *58*, 1216–1233.

(7) Karshikoff, A.; Ladenstein, R. *Trends Biochem. Sci.* **2001**, *26*, 550–556.

(8) Murphy, K. P.; Freire, E. *Adv. Protein Chem.* **1992**, *43*, 313–361.

(9) Montgomery, D.; Jordan, R.; McMacken, R.; Freire, E. *J. Mol. Biol.* **1993**, *232*, 680–692.

(10) Haynie, D. T.; Freire, E. *Anal. Biochem.* **1994**, *216*, 33−41.

(11) Xie, D.; Fox, R.; Freire, E. *Protein Sci.* **1994**, *3*, 2175–2184.

(12) Viguera, A. R.; Martinez, J. C.; Filimonov, V. V.; Mateo, P. L.; Serrano, L. *Biochemistry* **1994**, *33*, 2142–2150.

(13) McCrary, B. S.; Edmondson, S. P.; Shriver, J. W. *J. Mol. Biol.* **1996**, *264*, 784–805.

(14) Loladze, V. V.; Ermolenko, D. N.; Makhatadze, G. I. *J. Mol. Biol.* **2002**, *320*, 343–357.

(15) Bhat, T. N.; Bentley, G. A.; Boulot, G.; Greene, M. I.; Tello, D.; Dallacqua, W.; Souchon, H.; Schwarz, F. P.; Mariuzza, R. A.; Poljak, R. J. *Proc. Natl. Acad. Sci. U. S. A.* **1994**, *91*, 1089–1093.

(16) Martinez, J. C.; Filimonov, V. V.; Mateo, P. L.; Schreiber, G.; Fersht, A. R. *Biochemistry* **1995**, *34*, 5224–5233.

(17) Baker, B. M.; Murphy, K. P. *J. Mol. Biol.* **1997**, *268*, 557–569.

(18) Gonzalez, M.; Bagatolli, L. A.; Echabe, I.; Arrondo, J. L. R.; Argarana, C. E.; Cantor, C. R.; Fidelio, G. D. *J. Biol. Chem.* **1997**, *272*, 11288–11294.

(19) Frisch, C.; Schreiber, G.; Johnson, C. M.; Fersht, A. R. *J. Mol. Biol.* **1997**, *267*, 696–706.

(20) Xavier, K. A.; Shick, K. A.; SmithGill, S. J.; Willson, R. C. *Biophys. J.* **1997**, *73*, 2116–2125.

(21) Jelesarov, I.; Bosshard, H. R. *J. Mol. Recognit.* **1999**, *12*, 3–18.

(22) Spolar, R. S.; Livingstone, J. R.; Record, M. T. *Biochemistry* **1992**, *31*, 3947–3955.

(23) Myers, J. K.; Pace, C. N.; Scholtz, J. M. *Protein Sci.* **1995**, *4*, 2138–2148.

(24) Makhatadze, G. I.; Privalov, P. L. *Adv. Prot. Chem.* **1995**, *47*, 307–425.

(25) Robertson, A. D.; Murphy, K. P. *Chem. Rev.* **1997**, *97*, 1251–1267.

(26) Loladze, V. V.; Ermolenko, D. N.; Makhatadze, G. I. *Protein Sci.* **2001**, *10*, 1343–1352.

(27) Nozaki, Y.; Tanford, C. *J. Biol. Chem.* **1971**, *246*, 2211–2217.

(28) Eisenberg, D.; McLachlan, A. D. *Nature* **1986**, *319*, 199–203.

(29) Ooi, T.; Oobatake, M.; Nemethy, G.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U. S. A.* **1987**, *84*, 3086–3090.

(30) Makhatadze, G. I.; Privalov, P. L. *J. Mol. Biol.* **1990**, *213*, 375–384.

(31) Privalov, P. L.; Makhatadze, G. I. *J. Mol. Biol.* **1990**, *213*, 385–391.

(32) Murphy, K. P.; Privalov, P. L.; Gill, S. J. *Science* **1990**, *247*, 559–561.

(33) Murphy, K. P.; Gill, S. J. *J. Mol. Biol.* **1991**, *222*, 699–709.

(34) Privalov, P. L.; Makhatadze, G. I. *J. Mol. Biol.* **1992**, *224*, 715–723.

(35) Makhatadze, G. I.; Privalov, P. L. *J. Mol. Biol.* **1993**, *232*, 639–659.

(36) Privalov, P. L.; Makhatadze, G. I. *J. Mol. Biol.* **1993**, *232*, 660–679.

(37) Habermann, S. M.; Murphy, K. P. *Protein Sci.* **1996**, *5*, 1229–1239.

(38) Makhatadze, G. I.; Lopez, M. M.; Privalov, P. L. *Biophys. Chem.* **1997**, *64*, 93–101.

(39) Vangunsteren, W. F.; Berendsen, H. J. C. *Angew. Chem., Int. Ed. Engl.* **1990**, *29*, 992–1023.

(40) Van Gunsteren, W. F.; Bakowies, D.; Baron, R.; Chandrasekhar, I.; Christen, M.; Daura, X.; Gee, P.; Geerke, D. P.; Glattli, A.; Hunenberger, P. H.; Kastenholz, M. A.; Ostenbrink, C.; Schenk, M.; Trzesniak, D.; van der Vegt, N. F. A.; Yu, H. B. *Angew. Chem., Int. Ed.* **2006**, *45*, 4064–4092.

(41) Skipper, N. T. *Chem. Phys. Lett.* **1993**, *207*, 424–429.

(42) Skipper, N. T.; Bridgeman, C. H.; Buckingham, A. D.; Mancera, R. L. *Faraday Discuss.* **1996**, 141–150.

(43) Dang, L. X. *J. Chem. Phys.* **1994**, *100*, 9032–9034.

(44) Ludemann, S.; Schreiber, H.; Abseher, R.; Steinhauser, O. *J. Chem. Phys.* **1996**, *104*, 286–295.

(45) Ludemann, S.; Abseher, R.; Schreiber, H.; Steinhauser, O. *J. Am. Chem. Soc.* **1997**, *119*, 4206–4213.

(46) Mancera, R. L.; Buckingham, A. D. *Chem. Phys. Lett.* **1995**, *234*, 296–303.

(47) Mancera, R. L.; Buckingham, A. D.; Skipper, N. T. *J. Chem. Soc., Faraday Trans.* **1997**, *93*, 2263–2267.

(48) Rick, S. W.; Berne, B. J. *J. Phys. Chem. B* **1997**, *101*, 10488–10493.

(49) Rick, S. W. *J. Phys. Chem. B* **2000**, *104*, 6884–6888.

(50) Rick, S. W. *J. Phys. Chem. B* **2003**, *107*, 9853–9857.

(51) Shimizu, S.; Chan, H. S. *J. Chem. Phys.* **2000**, *113*, 4683–4700.

(52) Shimizu, S.; Chan, H. S. *J. Am. Chem. Soc.* **2001**, *123*, 2083–2084.

(53) Shimizu, S.; Chan, H. S. *Proteins: Struct., Funct., Genet.* **2002**, *48*, 15–30.

(54) Southall, N. T.; Dill, K. A. *Biophys. Chem.* **2002**, *101*, 295–307.

(55) Paschek, D. *J. Chem. Phys.* **2004**, *120*, 6674–6690.

(56) Paschek, D. *J. Chem. Phys.* **2004**, *120*, 10605–10617.

(57) Dill, K. A. *Biochemistry* **1990**, *29*, 7133–7155.

(58) Jorgensen, W. L.; Buckner, J. K.; Boudon, S.; Tiradorives, J. *J. Chem. Phys.* **1988**, *89*, 3742–3746.

(59) Nemethy, G.; Scheraga, H. A. *J. Phys. Chem.* **1962**, *66*, 1773–1789.

(60) Baldwin, R. L. *Proc. Natl. Acad. Sci. U. S. A.* **1986**, *83*, 8069–8072.

(61) Prabhu, N. V.; Sharp, K. A. *Annu. Rev. Phys. Chem.* **2005**, *56*, 521–548.

(62) Dang, L. X. *J. Chem. Phys.* **1992**, *97*, 1919–1921.

(63) Thomas, A. S.; Elcock, A. H. *J. Am. Chem. Soc.* **2004**, *126*, 2208–2214.

(64) Madan, B.; Sharp, K. *J. Phys. Chem.* **1996**, *100*, 7713–7721.

(65) Sharp, K. A.; Madan, B. *J. Phys. Chem. B* **1997**, *101*, 4343–4348.

(66) Sharp, K. A.; Madan, B.; Manas, E.; Vanderkooi, J. M. *J. Chem. Phys.* **2001**, *114*, 1791–1796.

(67) Gallagher, K.; Sharp, K. *Biophys. J.* **1998**, *75*, 769–776.

(68) Gallagher, K. R.; Sharp, K. A. *J. Am. Chem. Soc.* **2003**, *125*, 9853–9860.

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

(70) Mahoney, M. W.; Jorgensen, W. L. *J. Chem. Phys.* **2000**, *112*, b8910–8922.

(71) Berendsen, H. J. C.; Vanderspoel, D.; Vandrunen, R. *Comput. Phys. Commun.* **1995**, *91*, 43–56.

(72) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Model.* **2001**, *7*, 306–317.

(73) Thomas, A. S.; Elcock, A. H. *J. Am. Chem. Soc.* **2006**, *128*, 7796–7806.

(74) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.

(75) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

(76) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. *J. Comput. Chem.* **1997**, *18*, 1463–1472.

(77) Parrinello, M.; Rahman, A. *J. Appl. Phys.* **1981**, *52*, 7182–7190.

(78) Nose, S. *J. Chem. Phys.* **1984**, *81*, 511–519.

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

(80) Mahoney, M. W.; Jorgensen, W. L. *J. Chem. Phys.* **2001**, *114*, 363–366.

(81) Rocchia, W.; Alexov, E.; Honig, B. *J. Phys. Chem. B* **2001**, *105*, 6507–6514.

(82) Rocchia, W.; Sridharan, S.; Nicholls, A.; Alexov, E.; Chiabrera, A.; Honig, B. *J. Comput. Chem.* **2002**, *23*, 128–137.

(83) Lide, D. R. Properties of water in the range 0−100 °C. In *CRC Handbook of Chemistry and Physics*, 82nd ed.; Baysinger, G., Koetzle, T. F., Berger, L. I., Kuchitsu, K., Craig, N. C., Lin, C. C., Goldberg, R. N., Smith, A. L., Eds.; CRC Press LLC: Boca Raton, FL, 2001; pp 3−6.

(84) Elcock, A. H.; McCammon, J. A. *J. Phys. Chem. B* **1997**, *101*, 9624–9634.

(85) Elcock, A. H. *J. Mol. Biol.* **1998**, *284*, 489–502.

(86) Davis, M. E.; Madura, J. D.; Luty, B. A.; McCammon, J. A. *Comput. Phys. Commun.* **1991**, *62*, 187–197.

(87) Sitkoff, D.; Sharp, K. A.; Honig, B. *J. Phys. Chem.* **1994**, *98*, 1978–1988.

(88) Privalov, P. L.; Khechina, N. *J. Mol. Biol.* **1974**, *86*, 665–684.

(89) Privalov, P. L. *Adv. Protein Chem.* **1979**, *33*, 167–241.

(90) Privalov, P. L.; Gill, S. J. *Adv. Protein Chem.* **1988**, *39*, 191–234.

(91) Fu, L.; Freire, E. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 9335–9338.

(92) Baldwin, R. L.; Muller, N. *Proc. Natl. Acad. Sci. U. S. A.* **1992**, *89*, 7110–7113.

(93) Garde, S.; Hummer, G.; Garcia, A. E.; Paulaitis, M. E.; Pratt, L. R. *Phys. Rev. Lett.* **1996**, *77*, 4966–4968.

(94) Ashbaugh, H. S.; Truskett, T. M.; Debenedetti, P. G. *J. Chem. Phys.* **2002**, *116*, 2907–2921.

(95) Graziano, G.; Lee, B. *Biophys. Chem.* **2003**, *105*, 241–250.

(96) Cabani, S.; Gianni, P.; Mollica, V.; Lepori, L. *J. Solution Chem.* **1981**, *10*, 563–595.

(97) Bennaim, A.; Marcus, Y. *J. Chem. Phys.* **1984**, *81*, 2016–2027.

(98) Marcus, Y. *Biophys. Chem.* **1994**, *51*, 111–127.

(99) Plyasunov, A. V.; Shock, E. L. *Geochim. Cosmochim. Acta* **2000**, *64*, 439–468.

(100) Honig, B.; Nicholls, A. *Science* **1995**, *268*, 1144–1149.

(101) Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161–2200.

(102) Baker, N. A. *Curr. Opin. Struct. Biol.* **2005**, *15*, 137–143.

(103) Hendsch, Z. S.; Tidor, B. *Protein Sci.* **1994**, *3*, 211–226.

(104) Xiao, L.; Honig, B. *J. Mol. Biol.* **1999**, *289*, 1435–1444.

(105) Kumar, S.; Ma, B. Y.; Tsai, C. J.; Nussinov, R. *Proteins: Struct., Funct., Genet.* **2000**, *38*, 368–383.

(106) Zhou, H. X.; Dong, F. *Biophys. J.* **2003**, *84*, 2216–2222.

(107) Kumar, S.; Nussinov, R. *ChemBioChem* **2004**, *5*, 280–290.

(108) Danciulescu, C.; Ladenstein, R.; Nilsson, L. *Biochemistry* **2007**, *46*, 8537–8549.

(109) Fennell, C. J.; Gezelter, J. D. *J. Chem. Theory Comput.* **2005**, *1*, 662.

(110) Wick, C. A.; Siepmann, J. I.; Schure, M. R. *J. Phys. Chem. B* **2003**, *107*, 10623.

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

(112) Petsko, G. A. *Methods Enzymol.* **2001**, *334*, 469–478.

(113) Aghajari, N.; Van Petegem, F.; Villeret, V.; Chessa, J. P.; Gerday, C.; Haser, R.; Van Beeumen, J. *Proteins: Struct., Funct., Genet.* **2003**, *50*, 636–647.

(114) Arnorsdottir, J.; Kristjansson, M. M.; Ficner, R. *FEBS J.* **2005**, *272*, 832–845.

(115) Bae, E.; Phillips, G. N. *J. Biol. Chem.* **2004**, *279*, 28202–28208.

(116) Leiros, I.; Moe, E.; Lanes, O.; Smalas, A. O.; Willassen, N. P. *Acta Crystallogr., Sect. D* **2003**, *59*, 1357–1365.

(117) Violot, S.; Aghajari, N.; Czjzek, M.; Feller, G.; Sonan, G. K.; Gouet, P.; Gerday, C.; Haser, R.; Receveur-Brechot, V. *J. Mol. Biol.* **2005**, *348*, 1211–1224.

(118) Wang, E.; Koutsioulis, D.; Leiros, H. K. S.; Andersen, O. A.; Bouriotis, V.; Hough, E.; Heikinheimo, P. *J. Mol. Biol.* **2007**, *366*, 1318–1331.

(119) Myers, J. K.; Pace, C. N. *Biophys. J.* **1996**, *71*, 2033–2039.

(120) Gallivan, J. P.; Dougherty, D. A. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 9459–9464.

(121) Waters, M. L. *Biopolymers* **2004**, *76*, 435–445.

(122) *SigmaPlot*, version 10.0; Systat: Richmond, CA, 2006.

CT1000704

JCTC Journal of Chemical Theory and Computation

# Efficient Parallel Decomposition of Dynamical Sampling in Glass-Forming Materials Based on an "On the Fly" Definition of Metabasins

Dimitrios G. Tsalikis,[†] Nikolaos Lempesis,[†] Georgios C. Boulougouris,*[,†,‡,§,||] and Doros N. Theodorou*[,†]

*School of Chemical Engineering, National Technical University of Athens, Zografou Campus, GR-15780 Athens, Greece, Engineering Informatics and Telecommunications, University of Western Macedonia, Konstantinou Karamanli 55, GR-50100 Kozani, Greece, Department of Chemical Engineering, University of Patras, GR-26500 Patras, Greece, and Scienomics SARL, 17, Square Edouard VII, 75009 Paris*

Received August 12, 2009

**Abstract:** In this work, we propose a highly parallelizable sampling scheme designed for atomistic simulations of glassy materials in the vicinity of the glass-transition temperature $T_g$, based on the idea of inherent structures (IS). Glassy dynamics is envisioned as a combination of two types of motions: (a) an "in basin" vibrational motion in the vicinity of a potential energy minimum (IS), and (b) transitions from one basin to another. In order to perform efficient dynamical sampling in the vicinity of $T_g$, we propose an "on the fly" definition of metabasins (i.e., collections of basins communicating via fast transitions in which the system spends a sufficient time before moving on to a neighboring collection). Our criterion for defining metabasins is based on the rate of identification of new basins in the course of a canonical molecular dynamics (MD) run. In order to compute individual rate constants between basins and metabasins, we propose to follow a swarm of microcanonical MD trajectories initiated at phase-space points sampled by a canonical MD run that is artificially trapped within a metabasin. The execution time required by this highly parallelizable scheme is reduced dramatically, since no information exchange takes place between the microcanonical trajectories. Results from our parallel methodology are compared against results from artificially trapped canonical MD runs, in terms of the evaluated rate constants, and found to be in very good agreement. Parallel simulations have been conducted on up to 250 processors, achieving almost linear scaling. The validity of our definition of metabasins is confirmed by analysis of the resulting network of basins.

## Introduction

Glassy materials have assumed an important role in our life and consequently attract the interest of the scientific community in both applied and basic research. Over the years, glasses have been categorized according to various criteria: (a) based on the temperature dependence of dynamic viscosity,[1] into strong and fragile glass-forming liquids, and (b) based on the intramolecular interactions responsible for dynamical entrapment,[2] into repulsive and attractive glasses. Repulsive glasses are usually observed at high densities, where repulsive interactions become dominant, whereas attractive glasses appear due to a strong short-range attractive interaction.[3]

* Corresponding authors. E-mail: gboulougouris@uowm.gr (G.C.B.) and doros@central.ntua.gr (D.N.T.). Fax: +30 210 772 3112.

[†] National Technical University of Athens.

[‡] University of Western Macedonia.

[§] University of Patras.

[||] Scienomics SARL.

10.1021/ct9004245
Published on Web 03/19/2010

The research work described in this article is focused on studying the dynamics of glass-forming materials. Unfortunately, the broad range of time scales for molecular motion present in glassy systems poses severe limitations for molecular simulation in the vicinity of and below the glass-transition temperature $T_g$. Any discrete numerical solution of the time evolution equations of a microscopic model is bound by the time step of the discretization, which has to be smaller than the characteristic time of the fastest process present, thereby limiting our ability to track the time evolution out to the desired longest time scales. Therefore, brute force MD simulations are doomed to describe only a very short part of the spectrum of time scales characterizing motion in a glassy system. In order to address this problem, we will try to elucidate the dynamics of glassy materials in terms of their potential energy landscape.

One of the most important features of the potential energy landscape is the local minima of the energy, or "inherent structures" (ISs),[4,5] around which the system is expected to spend most of its time trapped, at least at low temperatures. Throughout this article, we will use the term "basin" to denote the set of configuration space points from which a steepest descent construction in the potential energy leads to a given IS.[4] The entire multidimensional configuration space can be tessellated into basins.

As in all molecular simulation methods, the size of the system is part of the simulation conditions and the macroscopic behavior can only be obtained as a thermodynamic limit. In the inherent structure approach, the condition that will minimize system size effects is the independence of cooperatively rearranging domains (in three-dimensional space) during the individual transitions from one basin to another. Assuming that the simulation system is sufficiently large, extensivity will result in independent transitions being executed by its various parts. For example, if one considers a model system of size double that of the original simulation system, one will observe elementary transitions, each of which involves a molecular rearrangement in only one of the two halves of the augmented system. Each of these rearrangements would have been detected as a single transition in an analysis of individual configurations of the original model system. Since each of the two sets of rearrangements (one set for each half of the doubled system) involves a different set of degrees of freedom, and the two sets of degrees of freedom are practically uncoupled, the rate constants for the individual rearrangements will be the same in the doubled system, as computed for the original system. Thus, the descriptions of dynamics on the basis of the original system and of the doubled system are equivalent, provided the original system is sufficiently large, in comparison to the size of "cooperatively rearranging domains". In this respect, basins of the original system would continue being relevant to the dynamics of the doubled system as well.

A prerequisite, of course, for this picture of extensivity to hold is that both the small and the large system configurations have been sampled from a distribution that is representative of the real glass under a given formation history. If this is the case, an inherent structure of the large model system will be essentially a combination of mutually independent inherent structures of smaller subsystems, into which the large system can be spatially decomposed. Consequently, the partition function of the large system trapped in the vicinity of one of its inherent structures will be essentially a product of partition functions of the subsystems, each trapped in the vicinity of its own inherent structure. It is this factorization that leads to rate constants for subsystem rearrangements being the same as computed for the individual subsystems and for the large system. This has been properly demonstrated in the work of Doliwa and Heuer,[6] where they examined finite-size effects in the same model. In their work they conclude that "a system of $N = 130$ particles behaves basically as two noninteracting systems of half the size."

Below $T_g$, the importance of ISs and basins to simulations of a variety of condensed matter systems has been extensively explored.[7−21] The "inherent structure picture" has been used as a tool to investigate and characterize the dynamics and the thermodynamics of atomistic systems in terms of their "landscape." Some of the most widely used concepts are the disconnectivity graphs[8−11,17,18,21] and the configurational entropy.[22−28] Another popular approach is the numerical validation, via simulations, of the theoretical predictions of the mode coupling theory and its extensions.[29−32] For extensive reviews on the subject, we refer the reader to the work of Heuer,[33] Sciortino,[22] Debenedetti, and Stillinger.[26]

It is worth noting that disconnectivity graphs have been used to visualize the energy landscape of the same model glass-former system studied here.[17−19] These studies have been conducted from both a thermodynamic perspective via the use of parallel tempering sampling[19] and from a dynamical perspective via extensive analysis of the local connectivity and the presence of "metabasins" in terms of the cage-breaking[17] process relevant to many transitions at the atomic level.[18]

Mapping long atomistic dynamics onto a discrete network of states has been used to analyze the folding pathways in protein-folding simulations.[34−41]

Most of the attempts described above are related with analyzing simulation results in terms of the IS picture. On the other hand, there are attempts to use IS ideas as a basis for accelerating the dynamical sampling. In view of the vast size and extensive nature of the potential energy landscape, it is crucial to sample the dynamical evolution between potential energy minima in an efficient manner. This has been demonstrated to be possible via the discrete path sampling (DPS) algorithm of Wales,[11,21] which is able to evaluate the most probable path of first-order transitions between potential energy minima linking two regions of the potential energy landscape, upon the assumption of an intermediate set of potential energy minima acting as the "activated" state.

One alternative attempt is the dynamical integration over a Markovian web (DIMW) methodology, developed by Boulougouris and Theodorou[13] in their effort to simulate the dynamics of an atomistic model of glassy atactic polystyrene over more than 10 orders of magnitude on the time scale at temperatures far below $T_g$. DIMW is also an alternative to kinetic Monte Carlo sampling, which enables the direct evaluation of the time-dependent probability of occupancy of states (here, basins) for a system undergoing successive transitions with first-order kinetics between basins in a landscape of infinite extent.

Although the DPS[11,21] and DIMW[13] algorithms can, in principle, be combined, their "philosophy" can be seen as complementary. DIMW creates a network of states through a breadth-first search that invokes no assumptions about ending and intermediate states, tracking a diffusion-like process in configuration space; DPS, on the other hand, may be seen as a depth-first search toward a known part of the landscape, upon the, most of the times logical, assumption of a rate-controlling intermediate region. Another important difference between the two methods is that the DIMW method aims at retrieving all dynamical information up to a certain time, that is the whole relaxation spectrum accessible to the potential energy landscape dynamics, whereas DPS, to our understanding, needs to include as target regions the ones that will potentially be relevant to a relaxation mechanism. The exhaustive dynamical information provided by DIMW may be of great importance. As discussed below, Boulougouris and Theodorou have recently shown[42] that knowledge of the spectrum of relaxation times, along with the eigenvectors of the dynamical transition matrix provided by the DIMW method, can be used in order to compute the time autocorrelation function and the dynamical relaxation spectrum of any observable; in this "EROPHILE" approach, a relaxation mode can be identified as a Euclidean vector in state probability space that decorrelates in a single exponential way. Thanks to the EROPHILE approach and the DIMW method, the creation of a complete basis set is guaranteed. DIMW can also be viewed as an extension to kinetic Monte Carlo of the integration over a Markovian web (IMW) method,[43] which allows the enrichment of ensemble averages and the combination of multiple integration levels. Based on DIMW, it is possible to construct an ever-expanding network of known or "explored" states, bounded by a set of "boundary" states, starting from an initial (small) set of states. For each explored state, all relevant transitions connecting it with its neighboring states have been located, and the corresponding transition rate constants computed by atomistic infrequent-event analysis. Boundary states are connected to explored states, but are not yet explored themselves. The time-dependent probability distribution among explored states is determined via analytical solution of the master equation for the "explored" states under absorbing boundary conditions for the "boundary" states. The set of explored states is expanded systematically whenever necessary, through a stochastic scheme that each time selects to include in the set of explored states a boundary state, according to the probability flux to that state. In their work Boulougouris and Theodorou[13] used multidimensional transition-state theory, within the harmonic approximation, in combination with a saddle-point search method[44−47] (specifically, the dimer method[46]) in order to locate and evaluate transitions and rate constants out of the states being explored on the fly. We note that the "dimer method" is actually an alternative implementation of hybrid eigenvector following, first described in ref 44.

Approaching $T_g$ from below, the number of "relevant" minima and saddle points increases dramatically. As a consequence, the computational cost for saddle-point calculations becomes prohibitively high. In order to overcome this obstacle, we have investigated the role of ISs in the vitrification process of glass-forming materials using a simple methodology, which is based on a combination of MD and potential energy minimization and on an extension of hazard plot analysis. This approach[48] showed that the dynamical transitions between basins can be described by a first-order kinetic scheme. More precisely, it was shown[49] that it is possible to reconstruct completely the dynamics of the atomistic system at a finite temperature, below $T_g$, based on the first-order kinetic network of interbasin transitions. This reconstruction corresponds to a "lifting"[50] of the coarse-grained Poisson process model of a succession of interbasin transitions to the detailed atomistic level. The excellent agreement obtained with full atomistic MD for temperatures around and below the glass transition temperature showed that an approach based on infrequent, uncorrelated transitions between basins is able to reproduce the full dynamics of the atomistic glassy system, where the Poisson approximation is valid.

The IS approach offers an additional advantage. The slow dynamics of the glassy systems described by the analytical solution for the time-dependent probability of occupancy of discrete states (basins around ISs) can be used for the identification of the molecular mechanisms that govern the dynamics. For this purpose, Boulougouris and Theodorou[42] developed a statistical mechanical-geometric formulation (EROPHILE) that expresses both state probabilities and all observables in the *same* Euclidean space, spanned by the eigenvectors of the symmetrized time evolution operator. EROPHILE is a general framework for computing the equilibrium and nonequilibrium behavior of systems evolving through a succession of transitions between discrete states. It provides a geometric representation of relaxation modes in a dual representation: (a) a mode corresponds to a perturbation from the equilibrium probability distribution among states that decays with time along a single exponential, and, most importantly, (b) a mode is identified with a linear combination of observables that, upon *any* perturbation from equilibrium, will return to equilibrium in a single exponentially decaying fashion. By applying EROPHILE to an atomistic model of a-PS, Boulougouris and Theodorou provided a molecular mechanism for the delta relaxation of a-PS, that of a net rotation of a single phenyl group around its stem.

As described above, well below $T_g$, the IS picture, coupled with infrequent event analysis, has been proven to be a reliable computational tool. Well above $T_g$ classical MD simulation is, in most cases, an efficient strategy. The vicinity of $T_g$, however, still remains a complex problem. Addressing this problem is very important since, as one expects, the temperature region in the vicinity of $T_g$ determines the quality,hence, the properties of the glass that we will obtain when we cool our glass-forming material far below $T_g$.[49] This happens for the simple reason that, in the temperature region far below $T_g$, the system practically "freezes" in the neighborhood of configurational space that it sampled just before the temperature dropped.

A necessary step for coarse graining the dynamics into the IS picture is the evaluation of rate constants for basin-to-basin transitions. This can be done with a variety of methods.[44−47,51−55] In the past we have used two distinct approaches for the rate calculations: a saddle-point search in combination with Fukui's intrinsic reaction coordinate (IRC) construction[56] and a harmonic approximation,[13] and MD simulations[48,49] in combination with hazard plot analysis.[57,58] The applicability of each approach depends on its computational demands. For temperatures far below $T_g$, an approach based on MD would suffer, since the system remains trapped in the vicinity of a handful of basins and does not escape even for times so long as to be inaccessible by classical MD, while a saddle-point search/IRC will show a much weaker dependence on barrier height and, therefore, will be preferable. On the other hand, for temperatures above $T_g$, saddle-point search suffers from the tremendous multitude of basins (several thousands in the course of nanoseconds for the model system sizes considered here) that need to be sampled, while brute force MD is expected to perform more efficiently. For the temperature range that is of primary interest here, in the vicinity of $T_g$, both methods suffer. The large number of visited basins makes the saddle-point search method computationally unaffordable, while classical MD must be pushed to its limits. The objective of this paper is to develop an efficient sampling method for this temperature range by achieving maximum parallelization. In a continuation of this work, we will show that it is even possible to accelerate the MD and sample rate constants over a very broad window of time scales with practically the same cost in real-time computation.

## Theory

In previous work,[48,49] the dynamics of a glassy material has been described by mapping onto a sequence of transitions between few basins, each basin constructed around an IS. As in that work, the system under study here was a mixture of Lennard-Jones (LJ) spheres that has been used widely to model glassy materials.[24,59,60] The mixture, initially proposed by Kob et al.,[60] consists of two different types of atoms, A and B, with atomic fractions 80% in A and 20% in B. The parameters of the model have been selected[60] in such a way that demixing is suppressed in order to suppress nucleation. Despite the fact that A atoms are larger than B atoms, they are assumed to have the same mass $m_A = m_B = 6.634 \times 10^{-26}$ kg. The LJ interaction parameters are $\varepsilon_{AA} = 1.65678 \times 10^{-21}$J, $\sigma_{AA} = 3.4 \times 10^{-10}$m, $\varepsilon_{BB} = 0.82839 \times 10^{-21}$J, $\sigma_{BB} = 2.992 \times 10^{-10}$m, $\varepsilon_{AB} = 2.48517 \times 10^{-21}$J, and $\sigma_{AB} = 2.72 \times 10^{-10}$m. The unit for reducing time is selected[59,60] as $[m_A\sigma_{AA}^2/(48\ \varepsilon_{AA})]^{1/2} = 3.10 \times 10^{-13}$ s, and the unit for temperature is $\varepsilon_{AA}/k_B = 120$ K. If the above LJ interaction parameters are reduced[61] by the values of the A−A interaction parameters, they read:[60] $\varepsilon_{AA} = 1.0$, $\sigma_{AA} = 1.0$, $\varepsilon_{BB} = 0.5$, $\sigma_{BB} = 0.88$, $\varepsilon_{AB} = 1.5$, and $\sigma_{AB} = 0.8$. In all calculations reported here, the molecular density of the system will be $1.1908\ \sigma_{AA}^{-3}$.

For this system in the supercooled state, Kob[59] and Shell et al.[24] have performed extensive studies, on the basis of which the mode coupling critical temperature $T_c$ is reported as 0.435 in reduced units (∼52.2 K).[59] For the same system, the glass transition temperature has been predicted[24] to be $T_g = 0.32$, that is to say, roughly equal to 38.4 K.

Most previous studies have focused on the region above $T_g$ close to the mode coupling[29−32] temperature $T_c$, where the system starts to deviate from ergodic sampling according to the mode coupling theory. It was shown that the number of basins visited per unit time by a $N = 641$ particle system depends strongly on the temperature of the system. For temperatures far below $T_g$, the system remains trapped in the vicinity of a handful of basins, even for times significantly longer than those accessible by conventional MD simulations (microseconds). As one increases the temperature approaching $T_g$, the number of basins sampled by traditional MD increases, and the temperatures close to and above $T_g$, it grows to several hundreds. For temperatures well above $T_g$ and $T_c$, MD sampling is sufficient to capture the basin-to-basin dynamics and reproduce the cage effect[59] and the process of atomic diffusion at long times.[49] Furthermore, it has been shown[48,49] that, for temperatures up to $T_g$, using artificially trapped MD simulations within each one of the visited basins and determining the transition rates out of the basins, it is possible fully to describe the system's atomistic dynamics in terms of both the coarse-grained motion from basin to basin and the intrabasin motion. The efficiency of such a procedure depends on the relation of the accessible simulation time to the time that the system needs to reach the basin boundary.

In this work, we introduce a self-consistent methodology that allows optimal use of MD over a wide range of temperatures. Here, by the term "optimal use", we refer to the ability of the method to automatically tune the length of MD trajectories used in order to sample inter- and intrametabasin transitions in an uncoupled fashion. For short times and low temperatures, the transitions between individual basins are rare events, while at higher temperatures, close to $T_g$, traditional MD can sample several basin-to-basin transitions, but the rare event is now the transition between collections of basins. Figure 1 shows results from a simulation in the NVT ensemble, which started from an equilibrated melt configuration (at 55K), under constant temperature close to $T_g$ ($T = 37$K). As one can see clearly in the inset, the system moves between three groups of basins (MB1, MB2, and MB3), where the transitions between groups are significantly slower in comparison to transitions between basins belonging to the same group. Following Heuer et al.,[62−64] we will refer to a collection of basins connected to each other through fast transitions as a "metabasin" (MB). Note that the reason why different MBs can be visually identified in Figure 1 is the "irreversible" nature (overall downhill direction of energy change) of the cooling process. Our target is to create a general autotuned method that enables the identification of a collection of basins and its characterization as a MB and that allows calculation of the transition rates from MB minima toward basins lying outside the MB boundary.

In the literature, several definitions have been proposed for the identification of a MB. Heuer[65] proposed an algorithm based on the IS trajectory, which can be summarized in the

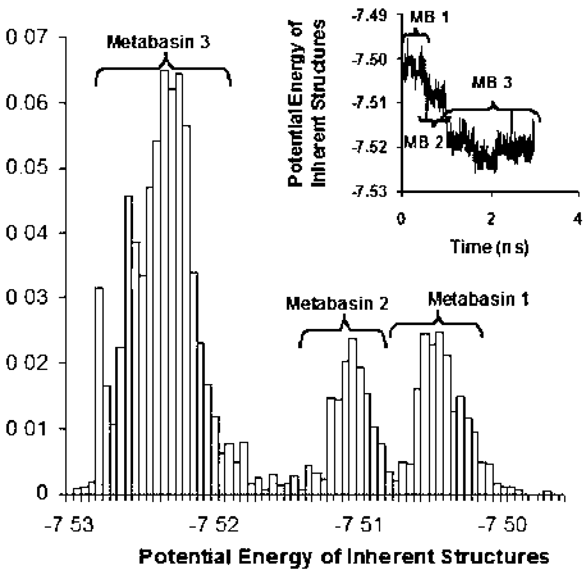


*Figure 1.* Distribution of the potential energies of inherent structures visited upon cooling the model system from $T =$ 55 to 37 K at a rate of 6 K/ns. In the inset to the diagram, the time evolution of the IS trajectory is given. The simulation time was 3 ns. Three metabasins can be identified visually.

following steps: (a) determine the time regions between the first and the last occurrence for each IS; and (b) group into a MB all basins for which there is an overlap in the corresponding time regions beyond a predefined time scale set to discard recrossing phenomena. Within this approach, the whole trajectory can be regarded, a posteriori, as a succession of different MBs. An alternative definition, independent of a specific trajectory, has been proposed[66,67] by Mauro and Loucks. Starting from rates between inherent structures, subsets are identified based on whether equilibration can be achieved within a prefixed time. As compared to the previous definitions, this allows one in some limit (no unbalanced transition rates)[66] to perform a partitioning of the configuration space into MBs, where the relaxation times within a MB are short compared to that of an observation time scale (prefixed time). However, in practice, many details of the potential energy landscape have to be discarded for the application of this approach. More details regarding MBs in glass-forming systems can be found in the review article of Heuer.[68] Recently,[18] the existence of MBs has been correlated with specific changes in the configuration space governed by the potential energy landscape, more precisely with the extent of cage-breaking (i.e., the molecular mechanism where the first neighbors of individual atoms are changing).

Another obstacle that one has to face when dealing with efficient sampling of basins and MBs is the existence of high-lying basins at the outskirts of the “most probable” basins that the system may visit. The reverse rate constant for leaving these basins to go back to the “probable” basins and MBs is much higher than the forward rate constant, and therefore, the actual probability of being in these peripheral basins is very small. On the one hand, one cannot discard such high-energy basins, since they may constitute passages to a different part of the landscape that may also be very “probable”, once it has been reached. On the other hand, one does not want to spend equal computational effort exploring the probable and improbable parts of the landscape. On the contrary, one would like to distribute the computational effort according to the probability of observing the system in each part of the landscape. Thus, besides the definition of MBs, an additional goal of our work is to implement the methodology in such a way that exploring states which do not belong to an important MB (i.e., which are occasionally visited by the system but are very quickly abandoned, as the system returns to a dominant MB) does not consume disproportionally large computational time.

A key feature of our methodology is the use of MD simulation itself in order to automatically tune the “dynamically accessible” part of the landscape under any conditions of temperature and density, given the available simulation time. In practice, we track the rate of exploring new minima by proper bookkeeping of the minima that have already been visited by our initial canonical MD run. When, for a given time interval, a plateau in the plot of the number of identified inherent structures versus simulation time is observed, which implies that the system configuration circulates within a confined collection of basins, we consider that the MB consists of the basins identified up to that point. In this way, we accomplish to group into a MB all the minima that are accessible from a starting minimum for a specific time window and to discriminate them from all other minima, for which sufficient sampling will require more computational effort. If our methodology is applied for low temperatures far below $T_g$, the MB is determined by a handful of minima or by even a single minimum (in the limiting case).

By construction, transitions from one MB toward its neighboring MBs will occur at a significantly longer time compared to the inner basin-to-basin transitions and to the simulation time used to define the MB. Therefore, the efficient sampling of transitions between MBs is, at least, an order of magnitude more demanding than sampling the inner MB. To accomplish such vigorous sampling, we developed an approach that allows the distribution of load within a parallel procedure that demands the same computational cost as the corresponding conventional MD run, but the results are obtained on a real-time scale more than two orders of magnitude faster. We propose a highly parallelizable scheme to achieve an efficient sampling of the MB dynamics. In this scheme, a long canonical MD trajectory is considered equivalent to a collection of microcanonical trajectories initiated at phase-space points sampled by a relatively short canonical MD simulation entrapped within the MB. Each microcanonical MD trajectory is terminated as soon as it exits the MB. The evaluation of the rate constants is based on hazard plot analysis of either the time difference between exiting and entering a basin, within the MB, or the measuring of the time that it takes the system to leave a basin (or the whole MB), given an equilibrated initialization within the same basin (or MB).

Hazard plot analysis is based on an evaluation of the cumulative hazard. The hazard rate, $h(t)$, is defined such that $h(t)dt$ is the probability that a system, which has survived a time $t$ since its last transition, will undergo a transition at a time between $t$ and $dt$. The cumulative hazard is defined as

$H(t) = \int_0^t h(t')dt'$. The probability that a transition occurs in time less than $t$ since the last transition is $P(t) = 1 - \exp[-H(t)]$. For a Poisson process, the hazard rate is constant $h(t) = \lambda$, the cumulative hazard is $H(t) = \lambda t$, and the probability is

$$P(t) = 1 - \exp[-\lambda t] \tag{1}$$

In our case of a Poisson process, the rate $\lambda$ can be extracted as the slope of a plot of the cumulative hazard $H$ versus the residence time $t$ at long times, when the effect of recrossing events has subsided, or as the negative slope of a plot of the quantity $\ln(1 - P^{\text{cum}}(t))$ versus the residence time. The last expression is based on solving eq 1 for $\lambda t$ and replacing $P(t)$ with its corresponding estimate $P^{\text{cum}}(t)$. The cumulative probability $P^{\text{cum}}(t_k)$ at a specific time $t_k$ can be determined as the ratio of the number of transitions that occurred with residence time up to $t_k$ divided by the total number of transitions. For a set of microcanonical trajectories, initiated at phase-space points sampled according to the Boltzmann weight that corresponds to the canonical ensemble, it is possible to group the transitions under the approximation that the system is at "*local*" *equilibrium.*

Consider that a set of transitions out of a given state with residence times less than or equal to $t_k$ is observed at $m$ energy levels ($E_1$, $E_2$, ..., $E_m$). We denote by $n_{E_i}$ the number of transitions that occurred in time less than $t_k$ under energy $E_i$. The number of transitions $k$ that occurred in time up to $t_k$ is $k = n_{E_1} + n_{E_2} + ... + n_{E_m}$. The total number of transitions out of the considered state is equal to $n$.

We can now estimate the cumulative probability as

$$P^{\text{cum}}_{\text{NVT}}(t_k) = \frac{k}{n} = \frac{n_{E_1} + n_{E_2} + ... + n_{E_m}}{n} = \frac{n_{E_1}}{n} + \frac{n_{E_2}}{n} + ... + \frac{n_{E_m}}{n} \tag{2}$$

Let $u_{E_i}$ be the number of transitions (at any residence time) out of the considered state observed under constant energy $E_i$. Then we can write eq 2 as

$$\begin{aligned} P^{\text{cum}}_{\text{NVT}}(t_k) &= \frac{n_{E_1}}{n} + \frac{n_{E_2}}{n} + ... + \frac{n_{E_m}}{n} = \frac{n_{E_1}}{n}\frac{u_{E_1}}{u_{E_1}} + \\ &\quad \frac{n_{E_2}}{n}\frac{u_{E_2}}{u_{E_2}} + ... + \frac{n_{E_m}}{n}\frac{u_{E_m}}{u_{E_m}}, \text{ or} \\ P^{\text{cum}}_{\text{NVT}}(t_k) &= \frac{n_{E_1}}{u_{E_1}}\frac{u_{E_1}}{n} + \frac{n_{E_2}}{u_{E_2}}\frac{u_{E_2}}{n} + ... + \frac{n_{E_m}}{u_{E_m}}\frac{u_{E_m}}{n} \end{aligned} \tag{3}$$

In our scheme, when the system is at local equilibrium, the terms $u_{E_i}/n$ will approximate the probabilities $p(E,T)$ to observe, in a canonical simulation under constant temperature $T$, the system at energy $E_i$.

Therefore, we can transform eq 3:

$$\hat{P}^{\text{cum}}_{\text{NVT}}(t_k) = \frac{n_{E_1}}{u_{E_1}}p^{\text{est}}(E_1,T) + \frac{n_{E_2}}{u_{E_2}}p^{\text{est}}(E_2,T) + ... + \frac{n_{E_m}}{u_{E_m}}p^{\text{est}}(E_m,T) \tag{4}$$

The term $n_{E_i}/u_{E_i}$ is the ratio of the number of transitions that occurred with residence time less than or equal to $t_k$ under constant energy $E_i$ during the simulation to the total number of transitions (at any residence time) under the same energy $E_i$. This ratio corresponds to the probability of observing a transition, at time less than or equal to $t_k$, if the simulation occurred under constant energy $E_i$ in the microcanonical ensemble. We replace in eq 4 all the terms $n_{E_i}/u_{E_i}$ with $P^{\text{cum}}_{NVE}(t_k)$:

$$\begin{aligned} P^{\text{cum}}_{\text{NVT}}(t_k) &= P^{\text{cum}}_{\text{NVE}_1}(t_k)p^{\text{est}}(E_1,T) + P^{\text{cum}}_{\text{NVE}_2}(t_k)p^{\text{est}}(E_2,T) + ...+ \\ &\quad P^{\text{cum}}_{\text{NVE}_m}(t_k)p^{\text{est}}(E_m,T), \text{ or} \\ P^{\text{cum}}_{\text{NVT}}(t_k) &= \int P^{\text{cum}}_{\text{NVE}}(t_k)p^{\text{est}}(E,T)\text{d}E \end{aligned} \tag{5}$$

where $\int p^{\text{est}}(E,T)\text{d}E = 1$. We have now expressed the probability to observe a transition in time $\le t_k$ at a specific temperature $T$ as an ensemble average over microcanonical trajectories initiated at phase-space points sampled in the course of an equilibrium canonical simulation. The initial points of each microcanonical trajectory are sampled based on the canonical ensemble in our algorithm.

Alternatively, one can derive eq 5 based on a superposition ansatz for the residence time distribution. Consider a system evolving along an NVT MD trajectory. We focus on transitions of the system into and out of a given state (basin or MB). The NVT trajectory will be assumed long enough to achieve local equilibration within a confined region of configuration space (MB or group of MBs, respectively) in which the system is temporarily trapped and which contains the considered state. In practice, the NVT MD trajectory is generated by coupling the system with a heat bath (e.g., through an extended ensemble technique). The time constant governing exchange of energy between the system and the heat bath must be long in comparison to the mean residence time in the state on which we focus; otherwise, our observations will be perturbed by interactions with the heat bath and will not reflect the true dynamics dictated by the potential energy hypersurface and the masses of system particles. Typically, each transition into and out of the state will primarily involve a relatively small subset of degrees of freedom of the system. The energy associated with this subset does fluctuate at a faster rate than the total energy of the system. Under these conditions, the total energy $E$ of the system between entry and immediately following the exit from the considered state will remain practically constant. By definition, then, the residence time distribution $\text{d}P/\text{d}t|_{\text{NVT}}$ determined in the course of a long NVT MD trajectory that allows the system to go in and out of the considered state can be related to the residence time distribution $\text{d}P/\text{d}t|_{\text{NVE}}$ that would be observed in the course of NVE MD trajectories conducted at energy levels $E$ as

$$\left.\frac{\mathrm{d}P}{\mathrm{d}t}\right|_{\mathrm{NVT}} = \int \left.\frac{\mathrm{d}P}{\mathrm{d}t}\right|_{\mathrm{NVE}} p(E,T)\mathrm{d}E \qquad (6)$$

Integrating with respect to time, this gives

$$P(t)|_{\mathrm{NVT}} = \int P(t)|_{\mathrm{NVE}}\, p(E,T)\mathrm{d}E \qquad (7)$$

In eqs 6 and 7, $p(E,T)$ is the probability of observing the system at energy $E$. The latter probability, however, is, by construction of the considered long, locally equilibrated NVT trajectory, proportional to the Boltzmann factor of the energy, retrieving eq 5.

Thus, we have expressed the cumulative probability to have undergone a transition at time $t$ at a specific temperature $T$ as an ensemble average of the corresponding cumulative probabilities calculated along microcanonical (NVE) trajectories. The initial phase-space points of the microcanonical trajectories are sampled by a canonical ensemble NVT MD simulation that has achieved local equilibration among a group of states to which the considered state belongs. The proposed approach can be envisioned as the reconstruction of an ensemble of NVT trajectories from a weighted ensemble of NVE trajectories. The correct dynamics can be sampled via either an NVE simulation or an NVT simulation in the limit where the thermostat interacts weakly with the system, in a way that does not perturb the system's time correlation functions. Nevertheless, our approach of using a swarm of NVE trajectories has a significant computational advantage; in the traditional NVT ensemble, once the interaction with the thermostat is weakened, the necessary time for thermal equilibration increases, whereas in our case, this is overcome by the proper weighting of each dynamical path. Despite the advantages of using NVE trajectories discussed above, we have actually tested whether it is possible to use NVT trajectories instead of NVE, and we have shown that, in practice, there is no significant difference.

**Molecular Simulation approach.** The first step in our methodology is to determine "on the fly" the local potential energy landscape that constitutes the MB, using a small-duration canonical MD simulation. Along an atomistic NVT MD simulation, at regular time intervals, the potential energy was minimized with the method of conjugate gradients[69] in order to identify the ISs of the MB. For the identification of ISs, Stillinger[4] proposed the use of the steepest descent method. In this work, we have chosen to use the method of conjugate gradients, which leads to the same IS as steepest descent in the overwhelming majority of cases but is significantly faster than steepest descent. To each one of the visited ISs we attribute an identity, storing its potential energy and its configuration. At this stage, we can have an estimate of the rate constants based on our previous work:[48,49] For each IS, we may choose to collect the transition times toward neighboring basins and their corresponding conditional probabilities and to compare the results of this stage with the final result of the proposed parallel scheme. The criterion we use to ascertain that the current MB has been sufficiently explored is based on the rate of identification of new (not already visited) basins of the potential energy and reflects the achievement of local equilibrium within the MB. Under these conditions, the system configuration circulates within a confined collection of basins, and a plateau in a plot of the number of identified ISs versus simulation time is observed. We consider that our MB consists of the number of basins identified up to that point. In Figure 2, we present a plot of the number of identified ISs versus the simulation time for a specific MB comprised of 290 minima.

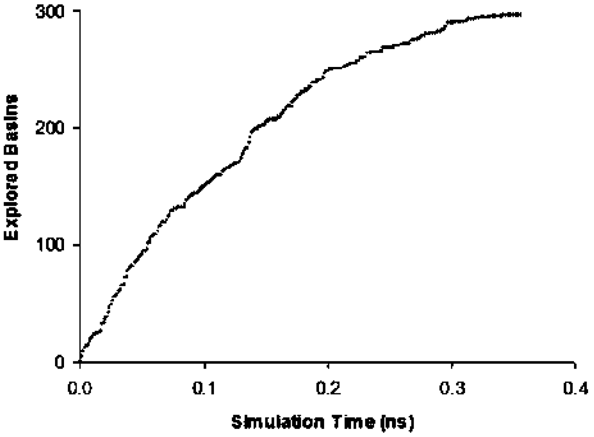


***Figure 2.*** Number of explored minima as a function of time at $T = 37$ K. When a plateau is observed for a prefixed time interval, the explored minima are considered to belong to the same MB.

This approach can easily be combined with the DIMW methodology for creating an ever-expanding network of MBs based on the following steps: (a) Define a MB based on short NVT MD runs, as described above; (b) Evaluate rate constants with the proposed methodology for all interbasin transitions within the MB and for basin-to-basin transitions terminating outside the MB; (c) Select an unexplored basin that lies outside of, but is connected to, the current (explored) MB based on the DIMW methodology; this is a starting point for identifying an *additional* MB; and (d) Loop back to step (a) to conduct an NVT MD out of the selected basin, but now, if the MD run reaches an explored basin of an explored MB, then go to step (b) and choose to leave the explored MB from a new unexplored basin (which can be one of the basins that are grouped as the additional MB). Continue the MD run until the rate of finding new MBs drops below a preset value. Note that, in this way, every new MB is not independent of the previous ones, but the union of identified MBs defines a set of basins in which the system will spend "sufficient" time before exploring new basins.

In order to evaluate the rate constants necessary for step (b), we proceed in the following manner: We produce an equilibrium sampling of phase-space points within the basins belonging to the MB via the execution of artificially trapped long canonical MD simulations. The artificial entrapment is implemented using reflective conditions at the MB boundaries. That is, once the system exits the MB, we invert the momenta (of the atoms and thermostat) stepping the system backward, returning it to the MB, following the procedure introduced in our previous work, where the inversion of the momentum was used to trap the system in individual basins.[48,49] The inversion of momenta is followed by an appropriate Gaussian randomization (that preserves the canonical distribution) of the momenta. This is used to perturb the system away from the original trajectory and assist the chaotic character of the system dynamics in sampling nearby trajectories. The new atomic momenta correspond to the imposed temperature of the canonical simulation, via the equipartition theorem, and sum to a total momentum of zero, as in the initial state. The duration of the artificially trapped simulation is ten times longer than that of the canonical MD that was used for the identification of the MB (step a). During the artificially trapped canonical simulation we store phase-space configurations of the system at constant time intervals. At the same time intervals, we minimize the potential energy of the system under constant volume and, thereby, make an assignment of sampled phase-space points to basins. Thereby, we can ensure that the stored phase-space point belongs to the sampled MB or record it as belonging to another MB. Within the simulation, we identify and store all transitions between basins belonging to the sampled MB and between basins of the sampled MB and unexplored basins lying outside the MB. It should be noted that the idea of a trapped simulation has been inspired by the novel methods pioneered by Voter[54,55] for studying the dynamics of rare events. Our aim is to combine these novel ideas with hazard plot analysis[58] for the calculation of rate constants via generation of a simple ensemble of MD trajectories in a high-dimensional energy landscape.

At this point, we propose a highly parallelizable scheme to achieve an efficient sampling of the MB dynamics *by considering the canonical molecular dynamics trajectory as equivalent to a collection of microcanonical trajectories* at total energies that have been chosen according to a Boltzmann weight corresponding to the canonical ensemble. To ensure that the initial states have been chosen with the appropriate Boltzmann weight, we confirm that the total energy distribution of the selected phase-state points provides a good estimate of the total energy distribution of the long canonical simulation along which they were sampled.

Out of each initial phase-space point, stored in the course of the "locally equilibrated" trapped long canonical MD simulation, we start two microcanonical (NVE) MD trajectories, one forward and one backward (by reversing the momentum of each particle). We continue our conjugate gradient potential energy minimizations at regular time intervals along each of these microcanonical trajectories, and we identify the minima based on the energy and on an Euclidean distance in configurational space (where atoms are considered distinguishable; swapping the identities of two particles results in a different basin). Each microcanonical MD simulation is terminated once the system visits an "unexplored" basin that does not belong to the current MB. We collect all the transitions between basins belonging to the sampled MB and between MB basins and unexplored basins lying outside the MB and perform hazard plot analysis on them as we have proposed for the basin-to-basin case.[48,49] We then determine the rate constant for each transition by ensemble averaging over the entire swarm of microcanonical trajectories according to eq 5.

This scheme of conducting a swarm of microcanonical MD simulations can be highly parallelized, since it does not require any communication between the simulations out of the different phase-space points. All necessary steps for the implementation of the parallel scheme are shown in Figure 3. In Figure 4, we provide a schematic representation of a configuration space, depicting the basic steps of the proposed methodology.

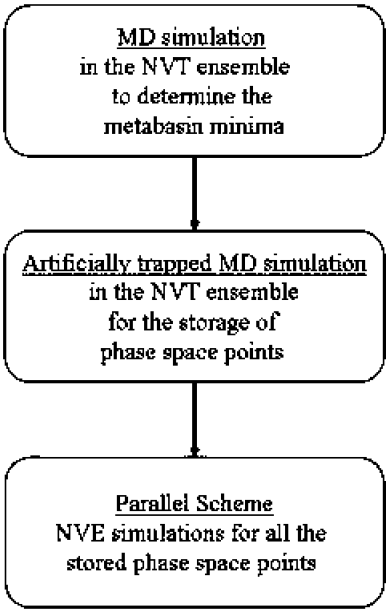


***Figure 3.*** Flow of calculations according to the proposed parallel scheme.

**Parallel Implementation.** The parallel computational work was conducted at the Supercomputing Center CINECA in Bologna, Italy. The simulations were performed on the IBM BCX/5120 cluster, which is mainly used for massively parallel applications and special high-end projects. The cluster consists of 2180 nodes, where each node is supplied with 2 Opteron dual core processors at 2.6 GHz. The nodes communicate via infiniband (5Gb/s) network. More information about the cluster can be found at http://www.cineca.it.

The implementation of the parallel scheme has employed a noncommercial MD code developed by the authors, using the Message Passing Inteface (MPI) library for distribution of the computational load. As we described above, we need to perform two microcanonical MD runs for each one of the stored phase-space points. Since these MD runs are completely independent, we use MPI to distribute the initial phase-space points among the available processors. Each processor performs a set of MD trajectories that start from the phase-space points assigned to it and end once the system has reached the "boundaries" of the MB. Energy minimization is performed, with the method of conjugate gradients,[69] at regular time intervals of the order of 0.1 ps. Each microcanonical MD simulation is completed when the system comes out of the MB, that is when, after the minimization procedure, the system is found in an "unexplored" IS that does not belong to the current MB. Thus, by construction, the simulation times of the independent microcanonical trajectories vary, and so does their corresponding execution time.

In order to share the computational cost into the parallel procedure, we developed two implementations: equally and unequally distributed configurations. In the first implementation, before the parallel simulation starts, we assign to the processors equal numbers of initial phase-space points. Each processor knows the number and identity of the phase-space

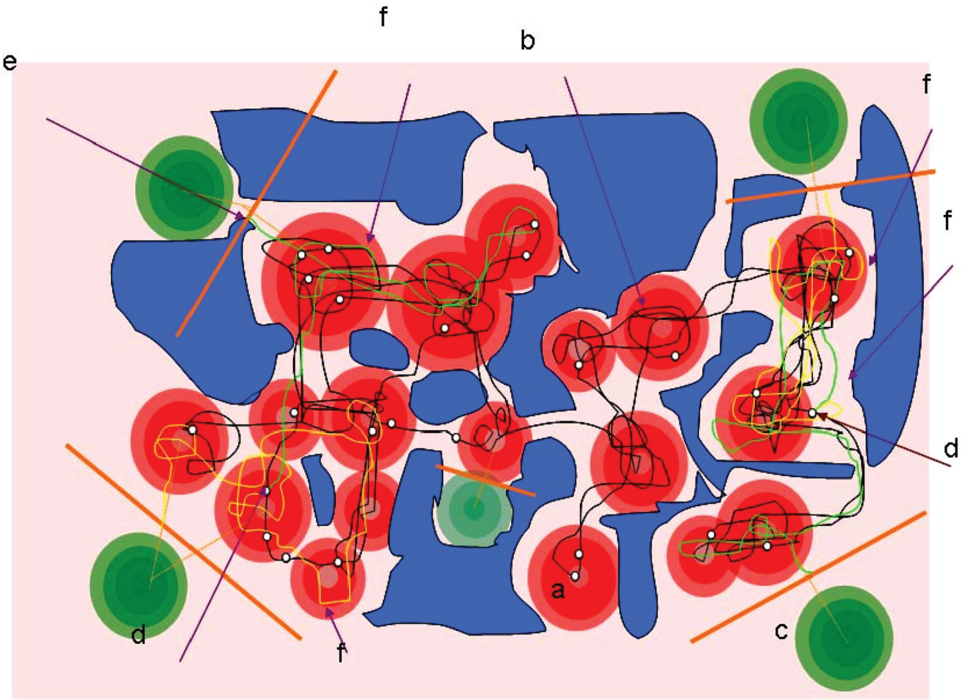


***Figure 4.*** A two-dimensional (2D) cartoon representation of a configuration space depicting the basic steps of the proposed methodology: (a) (Red regions) representation of inherent structures/basins which constitute the sampled MB. (b) (Black line) configuration space projection of the NVT MD trajectory used to define the MB. (c) (Green regions) neighboring basins, which are not part of the MB. (d) (Small white circles) configuration space projections of points in phase space sampled during an NVT MD trajectory (represented again by the black line) entrapped within the MB; these points are employed as the starting points for the swarm of NVE MD trajectories used to evaluate the rate constants. (e) Ending points of the NVE MD trajectories, i.e., points at which these trajectories leave the MB. (f) Two NVE MD trajectories (yellow and green) started from two of the stored phase-space points. For reasons of clarity, only one NVT MD trajectory is drawn here (black line). In reality we use two, one for the initialization of the MB and one for the collection of the NVE MD starting points. In blue, we depict the inaccessible (high potential energy) part of the configuration space; in reality, this constitutes a very large fraction of the configuration space.

points for which it will execute microcanonical simulations. This simplified parallel implementation was constructed conveniently, since the required programming cost is small. We proceeded to develop a more complex implementation, that of unequally distributed configurations, which aims at equalizing the distribution of computational load. Within this implementation, from the $N$ processors used, one (e.g., node0) is dedicated to “dealing” initial phase-space points out to the remaining $N - 1$ processors. Initially, node0 distributes, via the MPI library, one phase-space point to each processor. As soon as a processor (from the set of $N - 1$) completes the simulation of its assigned phase-space point, it communicates, via the MPI library, with node0, and a new phase-space point is assigned to it, until all phase-space points finish their microcanonical simulation.

## Results and Discussion

In order to validate the proposed scheme we compare, for a specific well-sampled basin of the identified MB, the transition rate obtained for exiting it from our proposed parallel methodology and from the artificially trapped canonical MD simulation in the MB. As one can clearly see in Figure 5, the resulting rates (long-time slopes of the cumulative hazard with respect to residence time) are in very good agreement. The comparison of Figure 5 has been performed for all the basins that constitute the MB.

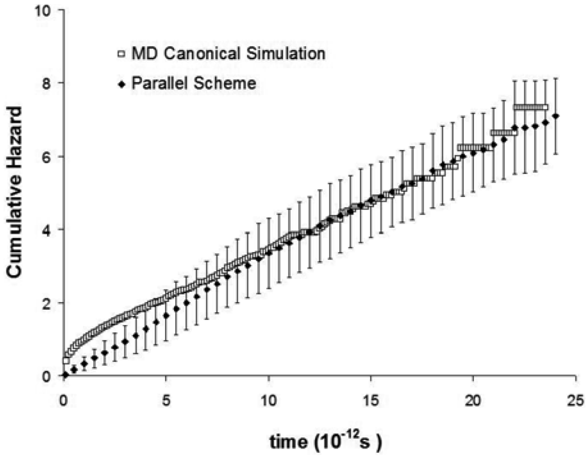


***Figure 5.*** Comparison between the results obtained from an artificially trapped trajectory in the canonical ensemble and the proposed parallel scheme for the calculation of the cumulative hazard to exit a specific basin of configuration space at temperature $T = 37$ K.

One of the advantages of using the hazard plot analysis is that, on top of the evaluation of the rate constant, the method validates whether the process is first order or not, since the linearity in the hazard plot is equivalent to an exponential distribution of the associated residence times. As has been described in our previous work, our hazard plot analysis has been designed to evaluate the sum of rates out of a basin or

***Table 1.*** Comparison of the Number of Saddle Points (Corresponding to Transitions between the MB minima) Identified with MD and with the Proposed Parallel Methodology[a]

| | saddle points | saddle points per execution time ($h^{-1}$) |
|---|---|---|
| MD | 3910 | 326 |
| parallel scheme | 24 271 | 1867 |

[a] Both applied within a MB. The MB consists of 290 minima.

***Table 2.*** Comparison of the Execution Time, the CPU Cost and the Simulation Time between MD and the Proposed Parallel Methodology[a]

| | execution time (h) | CPU cost (h) | simulation time (s) |
|---|---|---|---|
| MD | 12 | 12 | $3 \times 10^{-9}$ |
| parallel scheme | 12 + 1 | 250 | $7.7 \times 10^{-8}$ |

[a] The CPU cost for the parallel scheme is given by the product of the execution time of the "slowest" processor and the number of processors used.

a metabasin. Under the assumption of a Poisson process, each rate is the sum of the individual transition rates to any other basin or MB, through single or multiple routes. As has been described in our previews work,[48,49] basin-to-basin transitions are clearly of first order at low temperatures, whereas at temperatures higher than $T_g$, one has to look for MB-to-MB transitions to recover a clear first-order character.

Note that we have used two different but equivalent methods to determine the rate constants from our microcanonical MD trajectories based on hazard plot analysis. First, we have used the traditional idea proposed by Helfand[58] of analyzing the residence time within each of our discrete states (basins), i.e., the difference between the exit and entrance times. On the other hand, we also chose to analyze (again via hazard plot analysis) the ensemble of times that it takes to exit a basin (or a MB) when the initial phase-space point has been chosen according to the local equilibrium conditions within the state. We use the second hazard plot approach when we calculate the rate constants out of both the basin of the initial "stored" phase-space point and the MB itself toward the basins of neighboring MBs. Whereas, we use the first hazard plot approach when we calculate the rate constants between all other basins that we encounter after the initial one, until we reach the final basin of the trajectory. Our hazard plot calculations are performed along the canonical or the swarm of microcanonical MD trajectories.

In Table 1, we can see that using classical MD we observe only 3910 transitions between basins in the MB, while using the proposed parallel methodology this number becomes six times larger (24 271). In Table 2, we present a comparison of the computational cost and the simulation time between the classical MD and the parallel method. The simulation time of our parallel MD scheme is 20 times longer that that of the corresponding artificially trapped MD, and the total computational cost of the parallel scheme is also around 20 times larger.

To quantify the parallelizability of this scheme, we studied the speedup factor ($S_p$) and the parallel efficiency ($E_p$) as functions of the number of processors used. The speedup factor is defined by the following equation: $S_p = T_1/T_p$, where $p$ is the number of processors used, $T_1$ is the execution time of the sequential run, and $T_p$ is the execution time of the parallel scheme with $p$ processors. The parallel efficiency is defined as: $E_p = S_p/p$.[70,71]

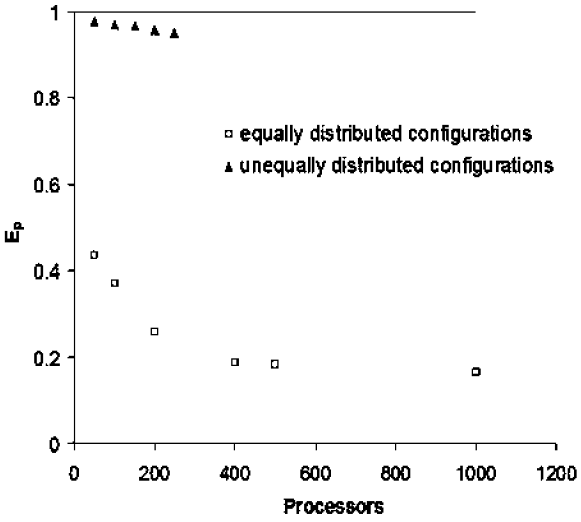


***Figure 6.*** Dependence of the efficiency factor ($E_p$) on the number of processors used for the implementation of equally distributed configurations (□) and for the implementation of unequally distributed configurations (▲).

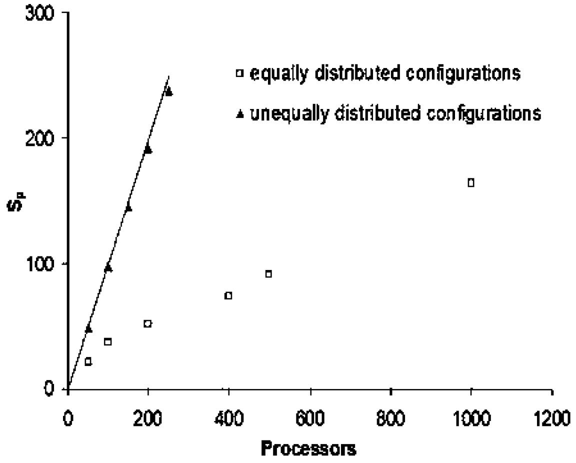


***Figure 7.*** Dependence of the speedup factor ($S_p$) on the number of processors used for the implementation of equally distributed configurations (□) and for the implementation of unequally distributed configurations (▲).

Results for the equally distributed configurations are presented in Figures 6 and 7 (□). As one can clearly see in these figures, this implementation for our proposed parallel application manages to reduce the execution time of our simulations even using up to 1000 processors. The speedup factor of this implementation indicates that using 1000 processors, with computational cost approximately six times larger than the corresponding sequential simulation ($E_{1000} \approx 0.164 \approx 1/6.11$), we shorten the execution time by a factor of 163 ($S_{1000} \approx 163$). The parallel efficiency obtained from this implementation is significantly lower than the optimum value that can be achieved by our parallel scheme. Since it does not demand any communication between the microca-

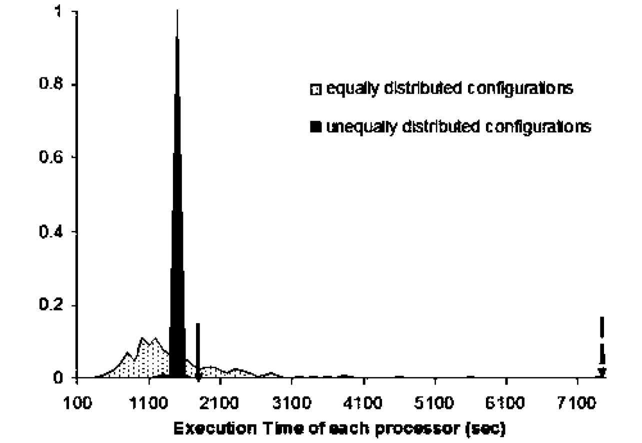


*Figure 8.* Representation of the distribution of the execution times of all processors for a simulation on 250 processors using the implementations of equally and unequally distributed configurations. The broken and solid arrows point at the maximum execution time for the first and second implementations, respectively.

nonical simulations, the proposed parallel scheme should allow higher values for the speedup factor (approaching linear dependence on the number of processors used − linear scaling) and for the parallel efficiency (approaching unity). The implementation of the equally distributed configurations suffers from load-balancing problems, due to the heterogeneity of the total simulation times of each processor. The distribution of execution times of the various processors for this implementation is presented in Figure 8 (gray area). Since completion of our parallel simulation demands that the simulation on the "slowest" processor be finished (computational cost is charged according to this rule in the supercomputing centers), the execution time of the overall simulation is determined by the execution time of the "slowest" processor. In order to solve this load-balancing problem, we developed the implementation of unequally distributed configurations. The dependence of the parallel efficiency and the speedup factor on the number of processors used, for this implementation, is presented in Figures 6 and 7, respectively. The distribution of execution times for a simulation on 250 processors, using the implementation of unequally distributed configurations, is presented in Figure 8. Using this implementation, as we can clearly see in Figure 6, we overcome the load-balancing problem plaguing the implementation of equally distributed configurations. As a result, we achieve an almost linear scaling for the speedup factor, even if we use a large number of processors ($S_{250} \approx 238$) and reduce the computational cost very significantly, close to the computational cost of the sequential simulation ($E_{250} \approx 0.92$). A great advantage of this implementation is that almost linear scaling is achieved by our parallel scheme, even if we use a large number of processors (250), independently of the simulated system size. For high-performance parallelization of MD simulations to be realized by domain decomposition into a large number of processors, most simulation software requires very big systems (of the order of $1 \times 10^5$ atoms) to be simulated. The implementation we propose here achieves its high performance even for very small systems, such as the system of Lennard-Jones spheres studied here (641 atoms).

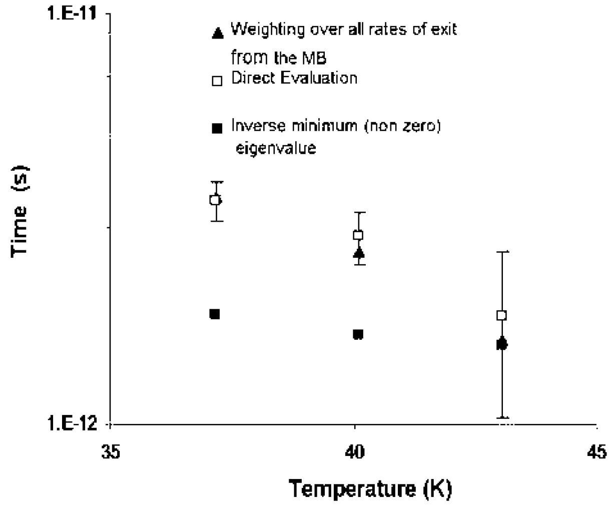


*Figure 9.* Comparison of the mean residence time in a MB that has been identified at 37 K, calculated by summing all rates for exiting the MB from any one of its basins, each weighted with the equilibrium probability of the corresponding basin (▲) and by hazard plot analysis over the artificially trapped microcanonical MD trajectories (□). The inverse minimum (nonzero) absolute eigenvalue of the rate-constant matrix for the specific MB is also shown, for comparison (■). The comparison has been performed for three temperatures (the basins constituting the MB are the same in all temperatures; what is changed is the temperature of the artificially trapped MD simulations).

The self-consistent methodology proposed for the definition of the MB aims at an automated selection of potential energy basins, among which local equilibration can be assumed to be established over the (long) time scales of interest. Equivalently, this can be translated into the requirement that the mean time spent in each of these minima is significantly longer than the time required for the achievement of local equilibration. For a system whose dynamics can be envisioned as a succession of transitions, with kinetics described by a first-order law, the response to any perturbation from the equilibrium probability distribution among states (in our case, basins) can be resolved into modes, each mode corresponding to a projection on an eigenvector of the matrix of transition-rate constants describing the dynamics of the system, appropriately symmetrized.[9,42,72,73] The matrix of transition-rate constants **K** is defined as follows:

$$K_{ij} = k_{j\to i} \forall i \neq j, \quad K_{ii} = -\sum_{j} k_{i\to j} \tag{8}$$

where $k_{i\to j}$ is the rate constant for the elementary transition from basins $i$ to $j$.

As in the "EROPHILE" approach, the $N$-dimensional vector $\mathbf{P}(t)$ of state probabilities for observing the system at time $t$ in each one of the $N$ distinct states $\mathbf{P}(t) \equiv (P_1(t), ..., P_i(t), ..., P_N(t))$ is transformed into a reduced vector $\tilde{\mathbf{P}}(t)$ with elements $\tilde{P}_i = P_i/\sqrt{P_i(\infty)}$. The elements of the transition-rate constant matrix are correspondingly transformed as $\tilde{K}_{ij} = K_{ij}\sqrt{P_j(\infty)}/\sqrt{P_i(\infty)}$. Under the condition of microscopic reversibility (detailed balance) on the rate constants ($k_{i\to j}P_i(\infty) = k_{j\to i}P_j(\infty)$) the matrix $\tilde{\mathbf{K}}$ is symmetric and similar to the transition-rate constant matrix $\mathbf{K}$. Since they are similar matrices, they have the same eigenvalues, which have to be all real due to the symmetry of the $\tilde{\mathbf{K}}$ matrix. On the other hand, the form of $\mathbf{K}$ guarantees that there is at least one 0 eigenvalue and that all other eigenvalues are negative. We denote these eigenvalues by $\lambda_0 = 0 \geq \lambda_1 \geq, ..., \geq \lambda_{N-1}$ and symbolize by $\tilde{\mathbf{u}}_n = (\tilde{u}_{1n}, \tilde{u}_{2n}, ..., \tilde{u}_{in}, ..., \tilde{u}_{Nn})$, the eigenvector of $\tilde{\mathbf{K}}$ corresponding to eigenvalue $\lambda_n$, $0 \leq n \leq N - 1$. The solution of the time-dependent state probabilities can be written as

$$P_i(t) = \sum_{n=0}^{N-1} \sum_{j=1}^{N} \frac{\sqrt{P_i(\infty)}}{\sqrt{P_j(\infty)}} \tilde{u}_{i,n}\tilde{u}_{j,n}e^{\lambda_n t}P_j(0), \text{ or}$$

$$\tilde{P}_i(t) = \sum_{n=0}^{N-1} \sum_{j=1}^{N} \tilde{u}_{i,n}\tilde{u}_{j,n}e^{\lambda_n t}\tilde{P}_j(0) \qquad (9)$$

or, in vector form:

$$\tilde{\mathbf{P}}(t) = \sum_{n=0}^{N-1} [\tilde{\mathbf{u}}_n \cdot \tilde{\mathbf{P}}(0)]\mathrm{e}^{\lambda_n t}\tilde{\mathbf{u}}_n = \tilde{\mathbf{P}}(\infty) + \sum_{n=1}^{N-1} [\tilde{\mathbf{u}}_n \cdot \tilde{\mathbf{P}}(0)]\mathrm{e}^{\lambda_n t}\tilde{\mathbf{u}}_n \qquad (10)$$

The eigenvectors $\tilde{\mathbf{u}}_n$ form an orthonormal basis set: $\tilde{\mathbf{u}}_m \cdot \tilde{\mathbf{u}}_n = \delta_{mn}$, $0 \leq m, n \leq N - 1$, with $\delta_{mn}$ being the Kronecker delta. They also satisfy $\sum_{n=0}^{N-1} \tilde{u}_{i,n}\tilde{u}_{j,n} = \delta_{ij}$.

Whereas eq 9 has been proposed in the past[9,73] in order to describe the time evolution of the state probabilities, in the work of Boulougouris and Theodorou, referred to as the "EROPHILE" approach,[42] the Euclidean orthonormal basis set created by the eigenvectors $\tilde{\mathbf{u}}_n$ is used for the first time, to our knowledge, to describe not only the state probabilities but also any real observables (i.e., their time-dependent averages and auto- and cross-correlations). For any observable $A$, it is possible to perform a transformation in the "EROPHILE" space, creating a Euclidean vector with components $\tilde{A}_i = A_i\sqrt{P_i(\infty)}$, where $A_i$ is the value of the observable in state $i$ of the system. The Euclidean vector $\tilde{\mathbf{A}}(t)$ can then be expressed in the same basis set as the state probability vector $\tilde{\mathbf{P}}(t)$.

EROPHILE is able to identify a relaxation mode either as a redistribution of the state probabilities in such a way that the return to equilibrium will occur along a single exponentially decaying function or, equivalently, as an observable for which the autocorrelation function will decay as a single exponential. From eq 10 it becomes obvious that each projection on every one of the eigenvectors evolves independently and, as the system approaches equilibrium, all mode contributions except the one corresponding to the 0 eigenvalue tend to 0 in an exponentially decaying fashion ($e^{\lambda_n t}$). Therefore, for times longer than the inverse minimum absolute value (nonzero) eigenvalue of the rate constant matrix, any perturbation of the system, no matter how big or improbable, will have been damped, and the system will have attained local equilibrium. To judge whether or not the system has achieved local equilibrium before leaving a given set of states, one has to compare the negative inverse of the smallest in absolute value nonzero eigenvalue of the transition rate matrix, $-1/\lambda_1$, with the average time it takes to leave the given set of states. In our case, we validate our "on the fly" identification of MBs by performing this comparison of $-1/\lambda_1$ against the mean residence time in the MB, as obtained from hazard plot analysis of our MD trajectories. As is shown in Figure 9, the former is three times smaller than the latter at 37 K and remains smaller at 40 and 43 K. An additional strong indication that local equilibration has been achieved within the MB is the equality between the values predicted for the mean residence time in the MB, as calculated directly via hazard plot analysis of the artificially trapped microcanonical MD trajectories and as estimated via summation of the rates for exiting the MB from any basin belonging to it, each weighted by the equilibrium probability of the basin, assuming local equilibrium (see Figure 9):

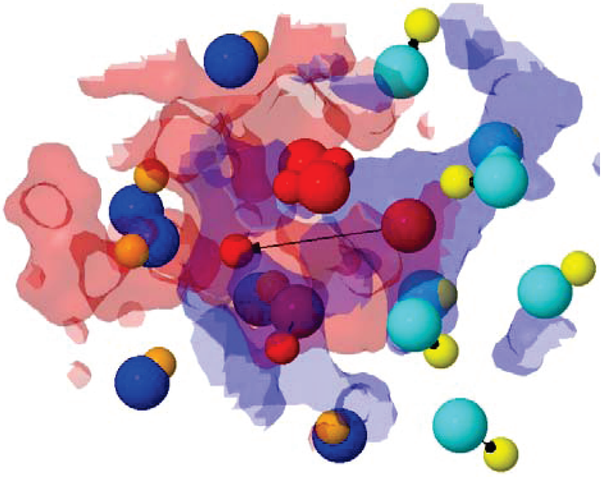


***Figure 10.*** Schematic representation of a cage-breaking event in a single jump from one potential energy minimum to a neighboring one. The positions of the atoms that participate in the transition are plotted with different sizes (initial: big; final: small), and vectors are drawn to indicate their displacements accompanying the transition. With red color we represent the atoms that remain first neighbors to the central atom experiencing the cage-breaking event, which is also shown in red. Cyan represents atoms that used to be first neighbors of the central atom but cease being so after the transition; their new positions are shown in yellow. Dark blue represents atoms which were not first neighbors of the central atom but come into its first coordination shell after the transition; their new positions are shown in orange. The blue surface depicts the volume accessible to the central molecule initially and the red finally, illustrating the cage change accompanying the transition.

$$k_{\mathrm{MB}\to\mathrm{out}} = \sum_i p_i^{\mathrm{eq}} k_{i\to j},\ i \in \mathrm{MB},\ j \notin \mathrm{MB} \qquad (11)$$

$$\langle t\rangle_{\mathrm{MB}\to\mathrm{out}} = 1/k_{\mathrm{MB}\to\mathrm{out}} \qquad (12)$$

Note that $p_i^{\mathrm{eq}}$ values are normalized to 1 in this calculation.

As mentioned above, this model system has been thoroughly studied in the past[17-20] and has provided very useful insights into the molecular motion relevant to relaxation in the vicinity of the glass transition, namely the "cage-breaking" process. More precisely, the change in the number of first neighbors accompanying a transition in the potential

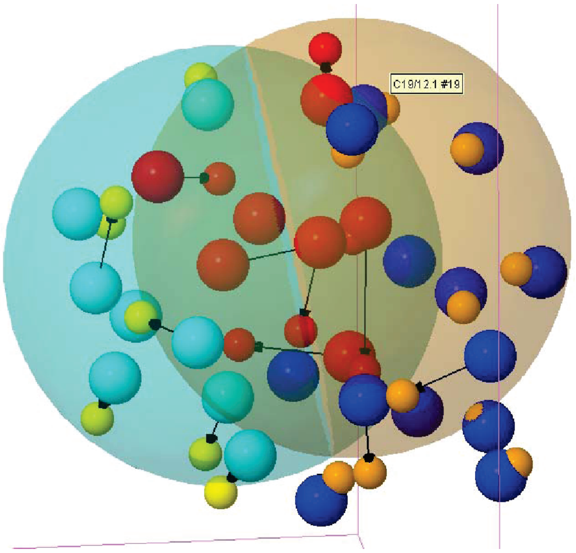


***Figure 11.*** Schematic representation of a more complex relaxation event that takes place in a single jump from one potential energy minimum to a neighboring one, belonging to a different MB. Coloring is as in Figure 10. We have also drawn cyan and light-brown spheres representing the first coordination shell, centered at the initial and final positions of the atom with the largest displacement. In this complex elementary move, atoms look like they are moving in a concerted way, exchanging their positions in a dance-like fashion.

energy landscape has been thoroughy examined in the work of Souza and Wales.[17] In our work, we also see cage-breaking events, as depicted in Figure 10, where a “central” atom jumps to a new cage after a single transition. On the other hand, when we investigate MB-to-MB transitions, we do not only see an enhancement of this process but also observe more complex relaxation mechanisms (Figure 11), wherein a number of atoms take each other’s positions moving in a, more or less, stringlike fashion, as if they were “dancing” in accordance with the “stringlike cooperative motion” demonstrated by the work[74] of Donati et al. In their work, by analyzing the van Hove correlation function produced via MD simulation for the same model system, Donati et al. showed that there is a fraction of mobile particles that at a characteristic time replace each other, executing a stringlike cooperative motion. Our result suggests that such a motion can be seen as a basin-to-basin transition intimately linked with the MB-to-MB transitions. We plan to investigate these more complex motions further in the future, since they require a great deal of cooperation between the atoms involved but may entail a less unfavorable energy barrier than the single cage-breaking events which move a molecule from one cage to another, bypassing (or pushing out) its first neighbors.

## Conclusions

We have developed an automated self-consistent method which can operate on the fly within a molecular dynamics (MD) simulation, allowing the identification of collections of basins and their characterization as metabasins (MBs). The criterion used to define MBs from short MD runs is based on the rate of identification of new, not already visited, basins. In practice, when for a given time interval we observe a plateau in the number of identified inherent structures (ISs) versus simulation time, implying that the system configuration circulates within a confined collection of basins, we consider that a MB has been identified, consisting of the basins visited up to that point. The proposed approach gives the ability to calculate the presence of a MB on the fly, the “minute” the system is trapped in a part of its configuration space; it does not require a postprocessing of the dynamics after the visit of several (at least two) MBs, as some previously proposed methods do.

The identification of a MB is followed by a calculation of the individual rate constants governing transitions between the basins constituting the MB and transitions toward basins that do not belong to the current MB. The computational cost for this calculation, which demands minimization of the potential energy for basin identification at regular time intervals, is significantly high, so we proceeded to develop a methodology that overcomes this obstacle. Our methodology distributes the vast computational cost associated with this calculation into a series of small duration-independent microcanonical MD simulations, conducted in parallel. Initial phase-space points for these microcanonical simulations are taken from a canonical MD simulation trapped within the MB. The execution time of the parallel microcanonical simulations is reduced dramatically, since our methodology does not require any information exchange between the simulations. Our results from the parallel methodology were compared against results from long artificially trapped canonical MD simulations and found to be in very good agreement. By implementing a scheme of unequally distributed configurations, wherein the parallel microcanonical MD runs are assigned to available processors in a manner that ensures good balancing of the computational load, we were able to achieve almost linear scaling ($E_{250} \approx 0.92$) on up to 250 processors, at a total computational cost similar to the cost of the corresponding sequential simulation. Additional advantages of our parallel methodology are that its applicability is independent of the system size (its high parallelization speedup and efficiency can be achieved even with very small systems, contrary to what happens with domain decomposition), and independent of the cluster architecture (shared memory or not). Using the proposed parallel methodology, we have examined the validity of our definition of the MB from the point of view of achievement of local equilibration among the basins that constitute the MB. To do so, we have compared the mean residence time between entry to and exit from the identified MB, calculated in two independent ways, with the time for decay of any perturbation away from local equilibrium within the MB. We observed that the mean residence time is significantly longer, indicating that local equilibrium has been achieved and that the MB has been successfully defined. The proposed parallel methodology distributes the vast computational cost of our calculation into practically independent runs, making it ideal as a backfill job on computing clusters. On the other hand, using a “workload management system” in scheduling the independent molecular dynamics runs has proved essential to efficiency. Beyond the usual benefits of parallel implementation, the Poisson character of our process (exponential distribution of the residence time) causes the individual independent runs to have widely varying computational cost.

This work is designed to extend the sampling ability of traditional MD simulation by utilizing an exremely efficient parallel approach. The approach is developed specifically to overcome some of the most vicious obstacles in the simulation of glassy systems, by turning them into an advantage. For example, the separation of time scales between intra- and intermetabasin transitions, which "immobilizes" traditional MD sampling, is now turned into an advantage, allowing for automated definition of fast and slow processes relative to the MD sampling ability. The novelty of the proposed approach lies in its design to overcome specific problems in simulating glass-forming systems. Furthermore, this is one of only few successful attempts to parallelize with high efficiency the calculation of dynamical properties. Last but not least, the use of the idea of a swarm of NVE trajectories to estimate the distribution of residence times in the NVT ensemble, and from that the rate constants, is, to our knowledge, novel and far from trivial.

The design of the algorithm aims at, as simple as possible, an implementation on the top of any existing MD package. The necessary tools are an MD simulator with the ability to perform minimizations; reflection and randomization of momentum[48,49] at specific intervals, depending on the result of the minimization; and a simple book-keeping procedure for visited minima. Furthermore, in all probability the proposed algorithm will be integrated into a general-purpose simulation package (probably as a tool in the MAPS program[75] of Scienomics SARL) based on a open-source MD platform.

**Acknowledgment.** This paper is part of the 03ED375 research project, implemented within the framework of the "Reinforcement Programme of Human Research Manpower" (PENED) and cofinanced by National and Community Funds (20% from the Greek Ministry of Development-General Secretariat of Research and Technology and 80% from E.U.-European Social Fund). Computational work was carried out under the HPC-EUROPA project (RII3-CT-2003-506079), with the support of the European Community—Research Infrastructure Action under the FP6 "Structuring the European Research Area" Program. The authors would like to thank Dr. Loukas Peristeras for his contribution to the development of the parallel code.

## References

(1) Angell, C. A. Structural instability and relaxation in liquid and glassy phases near the fragile liquid limit. *J. Non-Cryst. Solids* **1988**, *102*, 205–221.

(2) Dawson, K. A.; Foffi, G.; Sciortino, F.; Tartaglia, P.; Zaccarelli, E. Mode-coupling theory of colloids with short-range attractions. *J. Phys.: Condens. Matter* **2001**, *13*, 9113.

(3) Boulougouris, G. C.; Frenkel, D. Novel Monte Carlo scheme for systems with short-ranged interactions. *J. Chem. Phys.* **2005**, *122*, 244106.

(4) Stillinger, F. H.; Weber, T. A. Hidden structure in liquids. *Phys. Rev. A: At., Mol., Opt. Phys.* **1982**, *25*, 978–989.

(5) Theodorou, D. N.; Suter, U. W. Detailed molecular structure of a vinyl polymer glass. *Macromolecules* **1985**, *18*, 1467–1478.

(6) Doliwa, B.; Heuer, A. Finite-size effects in a supercooled liquid. *J. Phys.: Condens. Matter.* **2003**, *15*, S849–S858.

(7) Theodorou, D. N. In *Principles of molecular simulation of gas transport in polymers*; Yampolskii, Y., Pinnau, I., Freeman, B. D., Eds. John Wiley: Hoboken, NJ, 2006; pp 47−92.

(8) Wales, D. J.; Miller, M. A.; Walsh, T. R. Archetypal energy landscapes. *Nature* **1998**, *394*, 758–760.

(9) Becker, O.; Karplus, M. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *J. Chem. Phys.* **1997**, *106*, 1495–1517.

(10) Wales, D. J.; Doye, J. P. K.; Miller, M. A.; Mortenson, P. N.; Walsh, T. R. Energy landscapes: from clusters to biomolecules. *Adv. Chem. Phys.* **2000**, 115.

(11) Wales, D. J. Discrete path sampling. *Mol. Phys.* **2002**, *100*, 3285–3306.

(12) Wales, D. J. Calculating rate constants and committor probabilities for transition networks by graph transformation. *J. Chem. Phys.* **2009**, *130*, 204111.

(13) Boulougouris, G. C.; Theodorou, D. N. Dynamical integration of a Markovian web: A first passage time approach. *J. Chem. Phys.* **2007**, *127*, 084903.

(14) Jain, T. S.; de Pablo, J. J. Investigation of Transition States in Bulk and Freestanding Film Polymer Glasses. *Phys. Rev. Lett.* **2004**, *92*, 155505.

(15) Riggleman, R. A.; Douglas, J. F.; de Pablo, J. J. Characterization of the potential energy landscape of an antiplasticized polymer. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2007**, *76*, 011504.

(16) Papakonstantopoulos, G. J.; Riggleman, R. A.; Barrat, J. L.; de Pablo, J. J. Molecular plasticity of polymeric glasses in the elastic regime. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2008**, *77*, 041502.

(17) Souza, V. K. d.; Wales, D. J. Connectivity in the potential energy landscape for binary Lennard-Jones systems. *J. Chem. Phys.* **2009**, *130*, 194508.

(18) Souza, V. K. d.; Wales, D. J. Energy landscapes for diffusion: Analysis of cage-breaking processes. *J. Chem. Phys.* **2008**, *129*, 164507.

(19) Calvo, F.; Bogdan, T. V.; Souza, V. K. d.; Wales, D. J. Equilibrium density of states and thermodynamic properties of a model glass former. *J. Chem. Phys.* **2007**, *127*, 044508.

(20) Middleton, T. F.; Wales, D. J. Comparison of kinetic Monte Carlo and molecular dynamics simulations of diffusion in a model glass former. *J. Chem. Phys.* **2004**, *120*, 8134–8143.

(21) Wales, D. Some further applications of discrete path sampling to cluster isomerization. *Mol. Phys.* **2004**, *102*, 891–908.

(22) Sciortino, F. Potential energy landscape description of supercooled liquids and glasses. *J. Stat. Mech.: Theory Exp.* **2005**, P05015.

(23) La Nave, E.; Sastry, S.; Sciortino, F. Relation between local diffusivity and local inherent structures in the Kob-Andersen Lennard-Jones model. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2006**, *74*, 050501.

(24) Shell, M. S.; Debenedetti, P. G.; Panagiotopoulos, A. Z. A conformal solution theory for the energy landscape and glass transition of mixtures. *Fluid Phase Equilib.* **2006**, *241*, 147–154.

(25) Debenedetti, P. G.; Stillinger, F. H.; Shell, M. S. Model Energy Landscapes. *J. Phys. Chem. B* **2003**, *107*, 14434–14442.

(26) Debenedetti, P. G.; Stillinger, F. H. Supercooled liquids and the glass transition. *Nature* **2001**, *410*, 259–267.

(27) Crisanti, A.; Ritort, F. Inherent Structures, Configurational Entropy and Slow Glassy Dynamics. *J. Phys. Condens. Matter* **2002**, *14*, 1381–1395.

(28) Chowdhary, J.; Keyes, T. Energy Landscapes Composed of Continuous Intertwining Equipotential Ribbons. *J. Phys. Chem. B* **2004**, *108*, 19786–19798.

(29) Götze, W.; Sjögren, L. Relaxation processes in supercooled liquids. *Rep. Prog. Phys.* **1992**, *55*, 241–376.

(30) Götze, W.; Sjögren, L. The glass transition singularity. *Z. Phys. B: Condens. Matter* **1987**, *65*, 415.

(31) Götze, W. Aspects of structural glass transition. In *Liquids, Freezing and Glass Transition*; Hansen, J.-P., Levesque, D., Zinn-Justin, J., Eds. Elsevier Science Publishers: Amsterdam, The Netherlands, 1991; Vol. I, pp 287−503.

(32) Kob, W. Computer simulations of supercooled liquids and glasses. *J. Phys. Condens. Matter* **1999**, *11*, R85–R115.

(33) Heuer, A. Exploring the potential energy landscape of glass-forming systems: from inherent structures via metabasins to macroscopic transport. *J. Phys. Condens. Matter* **2008**, *20*, 373101.

(34) Voelz, V. A.; Bowman, G. R.; Beauchamp, K.; Pande, V. S. Molecular Simulation of ab Initio Protein Folding for a Millisecond Folder NTL9(1−39). *J. Am. Chem. Soc.* **2010**, *10*, 1021.

(35) Noe, F.; Schutte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 19011–19016.

(36) Noe, F. Probability distributions of molecular observables computed from Markov models. *J. Chem. Phys.* **2008**, *128*, 13.

(37) Snow, C. D.; Nguyen, N.; Pande, V. S.; Gruebele, M. Absolute comparison of simulated and experimental protein-folding dynamics. *Nature* **2002**, *420*, 102–106.

(38) Krivov, S. V.; Karplus, M. Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 14766–14770.

(39) Elmer, S. P.; Park, S.; Pande, V. S. Foldamer dynamics expressed via Markov state models. II. State space decomposition. *J. Chem. Phys.* **2005**, *123*, 7.

(40) Swope, W. C.; Pitera, J. W.; Suits, F. Describing protein folding kinetics by molecular dynamics simulations. 1. Theory. *J. Phys. Chem. B.* **2004**, *108*, 6571–6581.

(41) Swope, W. C.; Pitera, J. W.; Suits, F.; Pitman, M.; Eleftheriou, M.; Fitch, B. G.; Germain, R. S.; Rayshubski, A.; Ward, T. J. C.; Zhestkov, Y.; Zhou, R. Describing protein folding kinetics by molecular dynamics simulations. 2. Example applications to alanine dipeptide and beta-hairpin peptide. *J. Phys. Chem. B* **2004**, *108*, 6582–6594.

(42) Boulougouris, G. C.; Theodorou, D. N. Probing subglass relaxation in polymers via a geometric representation of probabilities, observables, and relaxation modes for discrete stochastic systems. *J. Chem. Phys.* **2009**, *130*, 044905–7.

(43) Boulougouris, G. C.; Frenkel, D. Monte Carlo Sampling of a Markov Web. *J. Chem. Theory Comput.* **2005**, *1*, 389–393.

(44) Munro, L. J.; Wales, D. J. Defect migration in crystalline silicon. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1999**, *59*, 3969–3980.

(45) Munro, L. J.; Wales, D. J. Rearrangements of bulk face-centred cubic nickel modelled by a Sutton-Chen potential. *Faraday Discuss.* **1997**, 409–423.

(46) Henkelman, G.; Jónsson, H. A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives. *J. Chem. Phys.* **1999**, *111*, 7010–7022.

(47) Barkema, G. T.; Mousseau, N. Event-Based Relaxation of Continuous Disordered Systems. *Phys. Rev. Lett.* **1996**, *77*, 4358.

(48) Tsalikis, D. G.; Lempesis, N.; Boulougouris, G. C.; Theodorou, D. N. On the role of inherent structures in glass-forming materials: I. The vitrification process. *J. Phys. Chem. B* **2008**, *112*, 10619–10627.

(49) Tsalikis, D. G.; Lempesis, N.; Boulougouris, G. C.; Theodorou, D. N. On the role of inherent structures in glass-forming materials: II. Reconstruction of the mean square displacement by rigorous lifting of the inherent structure dynamics. *J. Phys. Chem. B* **2008**, *112*, 10628–10636.

(50) Makeev, A. G.; Maroudas, D.; Panagiotopoulos, A. Z.; Kevrekidis, I. G. Coarse bifurcation analysis of kinetic Monte Carlo simulations: a lattice gas model with lateral interactions. *J. Chem. Phys.* **2002**, *117*, 8229–8240.

(51) Bolhuis, P. G.; Dellago, C.; Chandler, D. Sampling ensembles of deterministic transition pathways. *Faraday Discuss.* **1998**, *110*, 421–436.

(52) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. Transition Path Sampling: Throwing Ropes Over Rough Mountain Passes, in the Dark. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291–318.

(53) Voter, A. F.; Doll, J. D. Dynamical corrections to transition-state theory for multistate systems, Surface self-diffusion in the rare-event regime. *J. Chem. Phys.* **1985**, *82*, 80–92.

(54) Voter, A. Parallel replica method for dynamics of infrequent events. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1998**, *57*, R13985–R13988.

(55) Voter, A. Hyperdynamics: Accelerated molecular dynamics of infrequent events. *Phys. Rev. Lett.* **1997**, *78*, 3908–3911.

(56) Fukui, K. The Path of Chemical Reactions. The IRC Approach. *Acc. Chem. Res.* **1981**, *14*, 363–368.

(57) Helfand, E.; Wasserman, Z. R.; Weber, T. A. Brownian dynamics study of polymer conformational transitions. *Macromolecules* **1980**, *13*, 526–533.

(58) Helfand, E. Brownian dynamics study of transitions in a polymer chain ofbistable oscillators. *J. Chem. Phys.* **1978**, *69*, 1010–1018.

(59) Kob, W.; Andersen, H. C. *Phys. Rev. E: Stat. Phys., Plasmas, Fluids, Relat. Interdiscip. Top.* **1995**, *51*, 4626.

(60) Kob, W.; Andersen, H. C. Scaling behavior in the beta relaxation regime of a supercooled Lennard-Jones mixture. *Phys. Rev. Lett.* **1994**, *73*, 1376–1379.

(61) Allen, M. P.; Tildesley, D. J. Advanced Simulation techniques. In *Computer simulation of liquids*, 1sted.; Clarendon Press: Oxford, U.K., 1987; pp 212−219.

(62) Appignanesi, G. A.; Rodríguez, Fris J. A.; Montani, R. A.; Kob, W. Democratic Particle Motion for Metabasin Transitions in Simple Glass Formers. *Phys. Rev. Lett.* **2006**, *96*, 057801–057804.

(63) Doliwa, B.; Heuer, A. Hopping in a supercooled Lennard-Jones liquid: Metabasins, waiting time distribution, and diffusion. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2003**, *67*, 030501.

(64) Denny, R. A.; Reichman, D. R.; Bouchaud, J. P. Trap models and slow dynamics in supercooled liquids. *Phys. Rev. Lett.* **2003**, *90*, 025503.

(65) Buechner, S.; Heuer, A. Metastable states as a key to the dynamics of glass forming Liquids. *Phys. Rev. Lett.* **2000**, *84*, 2168–2171.

(66) Mauro, J. C.; Loucks, R. J.; Gupta, P. K. Metabasin approach for computing the master equation dynamics of systems with broken ergodicity. *J. Phys. Chem. A.* **2007**, *111*, 7957–7965.

(67) Mauro, J. C.; Loucks, R. J. Selenium glass transition: A model based on the enthalpy landscape approach and nonequilibrium statistical mechanics. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2007**, *76*, 174202.

(68) Heuer, A. Exploring the potential energy landscape of glass-forming systems: from inherent structures via metabasins to macroscopic transport. *J. Phys.: Condens. Matter* **2008**, *20*, 373101.

(69) Press, H. W.; Teukolsky, A. S.; Vettering, T. W.; Flannery, P. B. *Numerical Recipes: The Art of Scientific ComputingNumerical Recipes: The Art of Scientific Computing*, 3rd ed.; Cambridge University Press: Cambridge, U.K., 2007.

(70) Grama, A.; Gupta, A.; Karypis, G.; Kumar, V. *Introduction to Parallel Computing - Design and Analysis of AlgorithmsIntroduction to Parallel Computing - Design and Analysis of Algorithms*. 2nd ed.; The Benjamin/Cummings Publishing Company: Redwood City, CA, 2003.

(71) Fox, G. C.; Johnson, M. A.; Lyzenga, G. A.; Otto, S. W.; Salmon, J. K.; Walker, D. W. Solving problems on concurrent processors. In *General techniques and regular problems*, Prentice-Hall, Inc.: Upper Saddle River, NJ, 1988; Vol. 1, p 592.

(72) Berry, R. S.; Breitengraserkunz, R. Topography and dynamics of multidimensional interatomic potential surfaces. *Phys. Rev. Lett.* **1995**, *74*, 3951–3954.

(73) Berry, R.; Wales, D. Freezing, melting, spinodals, and clusters. *Phys. Rev. Lett.* 1989.

(74) Donati, C.; Douglas, J. F.; Kob, W.; Plimpton, S. J.; Poole, P. H.; Glotzer, S. C. Stringlike cooperative motion in a supercooled liquid. *Phys. Rev. Lett.* **1998**, *80*, 2338–2341.

(75) *Materials and Processes Simulations - MAPS*, 3.1 ed.; Scienomics: Paris, 2009.

CT9004245

JCTC Journal of Chemical Theory and Computation

# General Purpose Electrostatic Embedding Potential

Peter V. Sushko*,†,‡ and Igor V. Abarenkov§

*Department of Physics and Astronomy and London Centre for Nanotechnology, University College London, Gower Street, London WC1E 6BT, United Kingdom, WPI Advanced Institute for Materials Research, Tohoku University, Sendai 980-8577, Japan, and Department of Physics, St. Petersburg State University, St. Petersburg 198504, Russia*

Received September 10, 2009

**Abstract:** We present a method and a computer code for accurate calculation of electrostatic potential in an arbitrary crystalline lattice modeled using a finite system. The method is based on complementing a lattice unit cell with a set of point charges in order to annihilate simultaneously all components of *any number* of the lowest multipole moments. The positions and the values of the complementary charges are determined analytically. The electrostatic potential produced by each modified cell is short range, and the corresponding lattice series converges absolutely, which makes it convenient to use in embedded cluster calculations of solids, surfaces, and low-dimensional structures. The method is illustrated by application to the rutile $TiO_2$ and α-quartz $SiO_2$ lattices and to those of several complex minerals.

## 1. Introduction

In embedded cluster and quantum mechanics/molecular mechanics (QM/MM) methods, a QM description of a part of the system, called a “region of interest”, is combined with the empirical description of its surroundings. These methods are paticularly advantageous to use in cases where electronic states, associated with the region of interest, and those, associated with its environment, are separated in space and in energy. Numerous implementations of these techniques (see, for example, refs 1−14) and their applications to studies of large organic molecules,[15] solutions,[13,16,17] nanoparticles,[18] molecular crystals,[19,20] excitons[21,22] and reactions and properties of defects[23−27] have been reported.

The QM/MM methods often use a finite system (which will be referred to as a *nanocluster* or NC) to model the bulk, the surfaces, and the complex interfaces in crystals. It is well-known that the electrostatic potential (EP) inside a finite system depends on the choice of the structural element used to construct it. We illustrate this idea on the example of a nanocluster constructed as a 5 × 5 extension of a crystal unit cell, as shown schematically in Figure 1a. If the unit cell has zero dipole ($D$) and nonzero quadrupole ($Q$) moments, the electrostatic potential $V$ in the inner region of the nanocluster, indicated with a circle, depends on the details of the nanocluster structure. As the size of the nanocluster increases, the potential converges to the Ewald potential shifted by a constant, which can adopt any value from $-\infty$ to $+\infty$, depending on the shape of the nanocluster.

This creates problems if *absolute* positions of energy levels need to be calculated. In particular, the values of the ionization energies and the electron affinities, calculated for surface defects, depend on the EP in a nanocluster modeling this surface and, therefore, on the particular way the nanocluster is constructed. Accurate prediction of these properties is important for understanding a wide range of processes. For example, theoretical prediction of the MgO (001) surface ionization potential[5,28] has helped to map out the energy levels of surface oxygen vacancies, hydrogen defects, and nanoscale structural defects,[5,29,30] with respect to both the top of the surface valence band and the vacuum level. These theoretical results have been used successfully to develop mechanisms of complex photoinduced processes including charge transfer,[31] site-selective chemical reactions,[32] and atom desorption.[33,34] Recent similar results obtained for silica surfaces[25] may need to be reconsidered

[*] Corresponding author. E-mail: p.sushko@ucl.ac.uk.

[†] University College London.

[‡] Tohoku University.

[§] St. Petersburg State University.

10.1021/ct900480p © 2010 American Chemical Society
Published on Web 02/24/2010

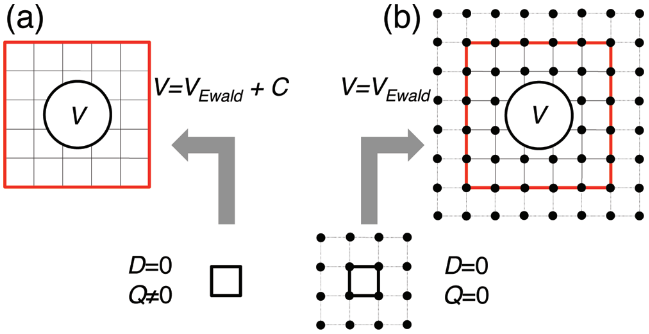


*Figure 1.* Electrostatic potential $V$ in the inner region of a nanocluster (top panels) depends on the lattice unit cell (bottom panels) used to generate it. (a) If the unit cell has zero dipole ($D$) and nonzero quadrupole ($Q$) moments, then $V(\boldsymbol{r})$ converges to $V_{\text{Ewald}}(\boldsymbol{r}) + C$ with increasing size of the nanocluster, where constant $C$ can adopt any value depending on the shape of the nanocluster. (b) If the unit cell is complemented with point charges so as both $D = 0$ and $Q = 0$, $V(\boldsymbol{r})$ converges to $V_{\text{Ewald}}(\boldsymbol{r})$ absolutely. The complementary charges (bold dots) are situated at the points equivalent to the unit cell corners. The boundaries of the nanocluster are indicated with bold solid lines.

because, unlike the MgO case, structural elements used to generate silica nanoclusters possessed nonzero quadrupole moments. The ambiguity in the values of the EP makes it difficult to set up a common reference for the bulk and surface defects in the same system and to compare results obtained for defects in different materials.

In this work, we suggest a method, together with a computer code,[35] which eliminates these problems. The method is based on complementing a lattice unit cell with point charges, which zero out *all multipole moments* of the cell up to any predefined $\mathscr{M}$ (see Figure 1b). If $\mathscr{M} \geq 2$, the potential in the inner region of the nanocluster converges absolutely to the result of the Ewald summation as the size of the nanocluster increases.[36] The complementary charges inside the nanocluster cancel each other out exactly by construction. Nonzero charges are situated only on the periphery of the nanocluster in a "skin" layer, thickness of which depends on the value of $\mathscr{M}$. Thus, the complementary charges provide corrective contribution to the EP inside the nanocluster without modifying the lattice structure in its inner region. This construction is convenient for embedded cluster calculations of crystalline systems and can be used to model bulk, surface, and low-dimensional structures.

Existing methods for constructing electrostatic embedding potential can be broadly divided into three categories:

**Grouping.** The crystal lattice can be divided into groups so as they have zero charge, dipole, and higher multipole moments.[37,38] The potential produced by each of these groups is short range, and the sum over the groups converges absolutely if the first nonzero moment is octupole. For example, the lattice building blocks for the rock-salt and perovskite (Figure 2a and b, respectively) structures can be selected so as their first nonzero multipole moments are $m = 6$ and 4, respectively.

**Fitting.** The difference between the EPs produced by a finite system and the corresponding infinite solid is fitted using a finite set of point charges.[8,39−41] This usually involves three steps: (i) calculating the Madelung potential using the Ewald summation method; (ii) introducing discretization in order to define the number of fitting charges and their positions; and (iii) solving a system of linear equations to find the best fit.

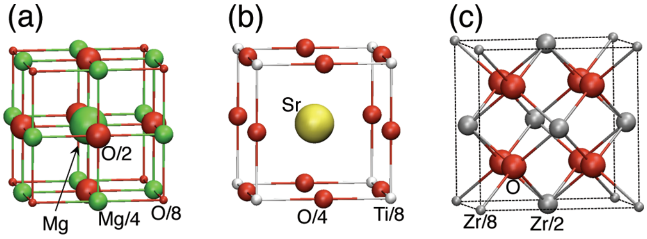


*Figure 2.* Examples of high-symmetry lattice building blocks: (a) MgO:$\text{Mg}(^1/_2\text{Mg})_{12}(^1/_2\text{O})_6(^1/_8\text{O})_8$; (b) $\text{SrTiO}_3$: $\text{Sr}(^1/_8\text{Ti})_8(^1/_4\text{O})_{12}$; (c) $\text{ZrO}_2$:$(^1/_2\text{Zr})_6(^1/_8\text{Zr})_8\text{O}_8$.

**Lattice Summation.** The difference between the Madelung and EPs due to a finite cluster can be also reproduced using multipoles associated with the lattice ions[42] or accounted for using the Ewald summation over infinite lattice.[14,43]

An implicit limitation of the fitting and lattice summation approaches is in their reference to the infinite lattice model and the periodic boundary conditions, which makes them difficult to apply to, for example, nonperiodic systems and irregular surfaces. In addition, the procedures related to the charge discretization and to operating with differences between the Madelung potential and the potential produced by a finite cluster are user dependent. In particular, the accuracy of the potential can vary depending on the complexity of the system, size, and shape of the nanocluster and on the choice of the number and location of the fitting charges.

The method suggested here regularizes the *grouping* approach. It neither requires calculating the EP of an infinite system nor involves fitting. Moreover, the electrostatic embedding potential is reproduced by a finite set of point charges, which makes it easy to employ standard codes for ab initio calculations.

## 2. Regularization of the Electrostatic Potential Series

Consider a crystalline lattice with lattice vectors $\boldsymbol{R}_k = k_1\boldsymbol{a}_1 + k_2\boldsymbol{a}_2 + k_3\boldsymbol{a}_3$, where $\boldsymbol{a}_1$, $\boldsymbol{a}_2$, and $\boldsymbol{a}_3$ are elementary translation vectors, and $k_1$, $k_2$, and $k_3$ are integers. If the position and the net charge of a $j$th atom in the unit cell are defined as $\rho_j$ and $e_j$, respectively, the EP for this lattice is given by the equation:

$$V(\boldsymbol{r}) = \sum_{kj} \frac{e_j}{|\boldsymbol{r} - \boldsymbol{R}_k - \rho_j|} \qquad (1)$$

where index $j$ runs through all atoms in the unit cell, and index $k$ runs through all unit cells of the system. In the case of infinite periodic lattices, the result of this summation depends on the order in which it is carried out.

The sum in eq 1 is not absolutely convergent, which means that, according to the Riemann series theorem, it can be made to converge to any desired value from $-\infty$ to $+\infty$ by a suitable rearrangement of the terms. Ewald[44] has proposed a summation procedure which regularizes eq 1 and transforms it into a sum of two absolutely converging series. This regularization is achieved by interchanging an integration and an infinite summation, which does not converge uniformly (see, for example, ref 45). Physical implication of this regularization are discussed elsewhere.[38]

An alternative approach, suggested by Madelung,[46] is based on regrouping the terms in eq 1:

$$V(\mathbf{r}) = \sum_k U(\mathbf{r} - \mathbf{R}_k) = \sum_k \left( \sum_{j=1}^{\mathcal{N}} \frac{e_j}{|\mathbf{r} - \mathbf{R}_k - \rho_j|} \right) \tag{2}$$

where $\mathcal{N}$ is the number of centers in the group, so as the infinite sum over $k$ becomes absolutely convergent. For this, it is necessary and sufficient to define the groups so as their zeroth, first, and second multipole moments, i.e., charge, dipole, and quadrupole, are all equal to zero simultaneously. The group of $\mathcal{N}$ centers does not have to coincide with the crystallographic unit cell but being translated with the corresponding lattice vectors should reproduce the whole infinite lattice.

Examples of such groups for MgO and $SrTiO_3$ are shown in Figure 2a and b, respectively. However, this method is difficult to generalize to complex lattices. It is straightforward to either select a unit cell or complement it with four fictitious charges,[47] so as to eliminate its dipole moment. However, eliminating the quadrupole and the higher electric moments in a general case is not trivial, and even in the case of cubic $ZrO_2$, the structural element shown in Figure 2c has nonzero components of the quadrupole moment.[48]

In the following, we consider a crystal unit cell which, being translated along its vectors, fills up the lattice without voids and overlaps. We demonstrate that for any, however complex, crystal such a cell can be complemented with a set of point charges so as: (i) all of its electric moments up to and including any finite $\mathcal{M}$ are eliminated simultaneously, and (ii) being translated with all possible lattice translation vectors, the modified unit cell reproduces the original lattice. The EP inside a finite system, formed of these groups, converges with the size of the system absolutely if $\mathcal{M} \geq 2$ and if the rate of the convergence is controlled by few well-defined numerical parameters.

Components of the $m$th multipole moment of a crystalline cell are defined as

$$Q_0(m_1, m_2, m_3) = \sum_{j=1}^{\mathcal{N}_0} e_j \rho_{jx}^{m_1} \rho_{jy}^{m_2} \rho_{jz}^{m_3} \tag{3}$$

where $\mathcal{N}_0$ is the number of atoms in the cell, and $m_1 + m_2 + m_3 = m$. For each $m$, sets $(m_1, m_2, m_3)$ can be represented as points with integer coordinates in the first octangle, as shown in Figure 3a.[36] The zero moment component $Q_0(0, 0, 0)$, i.e., the unit cell charge, corresponds to the point at the origin in Figure 3b. Components $Q_0(1, 0, 0)$, $Q_0(0, 1, 0)$, and $Q_0(0, 0, 1)$ of the first moment, i.e., the unit cell dipole, correspond to the three points shown in Figure 3c. Points corresponding to components of the quadrupole ($m = 2$) and octupole ($m = 3$) moments are shown in Figure 3 d and e, respectively.

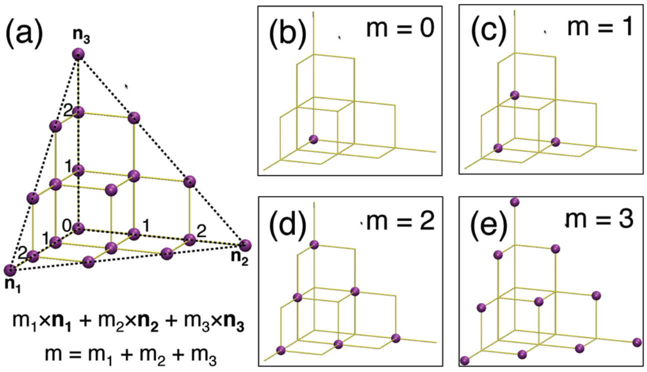


***Figure 3.*** Correspondence between points occupying sites with integer coordinates $m_1$, $m_2$, and $m_3$ and components of multipole moments $x^{m_1}y^{m_2}z^{m_3}$, where the moment $m$ is given by $m_1 + m_2 + m_3$.

In general, *all* components of *all* multipole moments up to $m = \mathcal{M}$ can be associated with integer-coordinate points inside a tetrahedron $T_{\mathcal{M}}$ with vertices at (0, 0, 0), ($\mathcal{M}$, 0, 0), (0, $\mathcal{M}$, 0), and (0, 0, $\mathcal{M}$).[36] Such tetrahedron for $\mathcal{M} = 3$ is shown with dashed lines in Figure 3a. For convenience, point (0,0,0) will be referred to as the main vertex of the tetrahedron.

We exploit this correspondence so as to eliminate multipole moments of the original unit cell. For that we introduce a set of point charges $e(\mathbf{n})$ at

$$\rho(\mathbf{n}) = n_1\mathbf{a}_1 + n_2\mathbf{a}_2 + n_3\mathbf{a}_3, \quad \mathbf{n} \in T_{\mathcal{M}} \tag{4}$$

where $\mathbf{n} = (n_1, n_2, n_3)$. Since the tetrahedron defined for components of multipole moments and the tetrahedron defined for charges $e(\mathbf{n})$ are equivalent, it is always possible to obtain such values of $e(\mathbf{n})$ that the multipole moments due to these charges cancel out the multipole moments of the original unit cell exactly:

$$\sum_{\mathbf{n} \in T_{\mathcal{M}}} e(\mathbf{n}) \rho_x^{m_1}(\mathbf{n}) \rho_y^{m_2}(\mathbf{n}) \rho_z^{m_3}(\mathbf{n}) = -Q_0(\mathbf{m}), \quad \mathbf{m} \in T_{\mathcal{M}} \tag{5}$$

In principle, this system of linear equations can be solved with respect to $e(\mathbf{n})$ using standard matrix diagonalization algorithms. However, the matrix elements can vary by several orders of magnitude, resulting in the loss of accuracy of the numerical solutions.

A general method for solving this system of equations *analytically* has been proposed in ref 36. Here we generalize this method in order to account for the fact that the EP function in a system has the same symmetry as the system itself. To this end, the tetrahedra $T_{\mathcal{M}}^{\alpha}$ ($\alpha = 1, 2, ..., 8$) and the corresponding set of extra charges are associated with *each* corner of the lattice unit cell.

For convenience, we will use fractional coordinates and assume that the corners of the unit cell are at the points ($\mp$1/2, $\mp$1/2, $\mp$1/2), where each combination of signs defines one of the corners and corresponds to one of the values of $\alpha$. Then, positions of the main vertices of the tetrahedra $T_{\mathcal{M}}^{\alpha}$ are given by

$$\boldsymbol{r}_0^\alpha = \mp\left(\frac{1}{2} + v\right)\boldsymbol{a}_1 \mp \left(\frac{1}{2} + v\right)\boldsymbol{a}_2 \mp \left(\frac{1}{2} + v\right)\boldsymbol{a}_3 \qquad (6)$$

where parameter $v$ defines the shift of $T_{\mathscr{M}}^\alpha$ with respect to the corresponding corner of the unit cell, and the charges $e^\alpha(\boldsymbol{n})$ are positioned at (compare with eq 4)

$$\rho^\alpha(\boldsymbol{n}) = \boldsymbol{r}_0^\alpha \pm n_1\boldsymbol{a}_1 \pm n_2\boldsymbol{a}_2 \pm n_3\boldsymbol{a}_3 \qquad (7)$$

The values of charges $e^\alpha(\boldsymbol{n})$ for each of the tetrahedra can be calculated separately in order to eliminate a fraction of the multipole moments of the original unit cell. For simplicity, for each α we choose, the coordinate system in which $\boldsymbol{r}_0^\alpha = 0$, and require that the extra charges associated with a tetrahedron $T_{\mathscr{M}}^\alpha$ cancel 1/8th of the unit cell multipole moments, i.e., $P^\alpha(\boldsymbol{m}) = (1/8)P_0(\boldsymbol{m})$, where $P_0(\boldsymbol{m})$ defines multipole moments of the original unit cell calculated in the same coordinate system and expressed in fractional coordinates. Then, eq 5 becomes

$$\sum_{\boldsymbol{n}\in T_{\mathscr{M}}^\alpha} e^\alpha(\boldsymbol{n})(\pm n_1)^{m_1}(\pm n_2)^{m_2}(\pm n_3)^{m_3} = -P^\alpha(\boldsymbol{m}), \quad \boldsymbol{m} \in T_{\mathscr{M}}^\alpha \qquad (8)$$

where the signs are defined by the orientation of $T_{\mathscr{M}}^\alpha$ and are opposite to those in eq 6.

This system of equations can be solved analytically with the help of auxiliary functions:[36]

$$G_k(x) = \begin{cases} 0 & \text{if } k = 0 \\ \prod_{j=0}^{k-1}(x - j) & \text{if } k > 0 \end{cases} \qquad (9)$$

which, for integer values of the argument, become

$$G_k(n) = \begin{cases} 0 & \text{if } n < k \\ \dfrac{n!}{(n-k)!} & \text{if } n \geq k \end{cases} \qquad (10)$$

Functions $G_k(x)$ are polynomials of $x$:

$$G_k(x) = \sum_{m=0}^{k} g(k, m)x^m \qquad (11)$$

and their coefficients $g(k, m)$ can be calculated analytically using recurrence equations:

$$g(k + 1, 0) = kg(k, 0) = g(1, 0) = 0$$
$$g(k + 1, k + 1) = g(k, k) = g(0, 0) = 1$$
$$g(k + 1, m) = g(k, m - 1) - kg(k, m), \quad m = 1, ..., k \qquad (12)$$

In one-dimensional case, the system of eqs 8 transforms into

$$\sum_{n=0}^{\mathscr{M}} (\pm n)^m e^\pm(n) = -P^\pm(m), \quad m = 0, 1, ..., \mathscr{M} \qquad (13)$$

where we adopted the ± sign instead of α (α = 1, 2). Multiplying these equations by $(\pm 1)^m g(k, m)$ and summing over $m$ we obtain

$$\sum_{m=0}^{k}\sum_{n=0}^{\mathscr{M}} g(k, m)n^m e^\pm(n) = -\sum_{m=0}^{k}(\pm 1)^m g(k, m)P^\pm(m) \qquad (14)$$

where $k = 0, ..., \mathscr{M}$. The right side of this equation is known and, for brevity, is denoted as $f^\pm(k)$, while the left side contains the expansion of $G_k(n)$ over powers of $n$:

$$\sum_{n=0}^{\mathscr{M}} G_k(n)e^\pm(n) = f^\pm(k), \quad k = 0, ..., \mathscr{M} \qquad (15)$$

Taking into account the properties of $G_k(n)$ (see eq 10), we obtain backward relations for the charges $e(n)$:

$$e^\pm(\mathscr{M}) = \frac{1}{\mathscr{M}!}f^\pm(\mathscr{M}) \qquad (16)$$

and

$$e^\pm(k) = \frac{1}{k!}\left(f^\pm(k) - \sum_{n=k+1}^{\mathscr{M}} \frac{n!}{(n-k)!}e^\pm(n)\right) \qquad (17)$$

where $k = \mathscr{M} - 1, \mathscr{M} - 2, ..., 0$.

Similar, although more complex, backward recurrence relations can be obtained for the three-dimensional case[36] and for any tetrahedron $T_{\mathscr{M}}^\alpha$. For simplicity, we omit the explicit index α, since the reference to each tetrahedron is incorporated in the choice of signs for $(\pm 1)^{m_i}$ ($i = 1, 2, 3$) in eq 8 and the values of the multipole moments $P(\boldsymbol{m})$, calculated in the coordinate system selected for each tetrahedron, as described above.

We renumerate points $(k_1, k_2, k_3)$ of a tetrahedron $T_{\mathscr{M}}$ from 1 to $N_{\mathscr{M}}$, using a single index $k = \varphi(k_1, k_2, k_3)$ and the explicit relation (see ref 36 for details):

$$k = \varphi(k_1, k_2, k_3) =$$
$$\frac{1}{6}k_1[3\mathscr{M}^2 + 12\mathscr{M} + 11 - 3(\mathscr{M} + 2)k_1 + k_1^2]$$
$$+k_2(\mathscr{M} + 1 - k_1) - \frac{1}{2}k_2(k_2 - 1) + k_3 + 1 \qquad (18)$$

In this numeration, the first index $k = 1$ corresponds to the point (0,0,0) and the last index $k = N_{\mathscr{M}}$ corresponds to the point ($\mathscr{M}$,0,0) of the tetrahedron.

Using the single index numeration, we introduce functions:

$$\tilde{e}(n) = e(n_1, n_2, n_3)$$
$$\tilde{G}_k(n) = G_{k_1}(n_1)G_{k_2}(n_2)G_{k_3}(n_3)$$

and

$$\tilde{f}(k) = f^{\pm\pm\pm}(k_1, k_2, k_3) \qquad (19)$$

where $\tilde{f}(k)$ is known

$$\tilde{f}(k) = -\sum_{m_1=0}^{k_1}\sum_{m_2=0}^{k_2}\sum_{m_3=0}^{k_3} P^\alpha(\boldsymbol{m})(\pm 1)^{m_1}(\pm 1)^{m_2}(\pm 1)^{m_3}$$
$$\times g(k_1, m_1)g(k_2, m_2)g(k_3, m_3) \qquad (20)$$

and the choice of signs is defined by that in eq 8. Then, the recurrent relations analogous to eqs 16 and 17 can be written as

$$\tilde{e}(N_{\mathscr{M}}) = \frac{1}{\tilde{G}_{N+_{\mathscr{M}}}(N_{\mathscr{M}})}\tilde{f}(N_{\mathscr{M}})$$

$$\tilde{e}(k) = \frac{1}{\tilde{G}_k(k)}\left(\tilde{f}(k) - \sum_{n=k+1}^{N_{\mathscr{M}}} \tilde{G}_k(n)\tilde{e}(n)\right) \tag{21}$$

where $k = N_{\mathscr{M}} - 1,\ N_{\mathscr{M}} - 2,\ ...,\ 0$, and

$$\tilde{G}_k(n) = \frac{n_1!}{(n_1 - k_1)!}\frac{n_2!}{(n_2 - k_2)!}\frac{n_3!}{(n_3 - k_3)!}, \quad k \le n \tag{22}$$

We note that for nanoclusters with characteristic size of more than $2\mathscr{M}$ unit cells, there is a finite inner region in which all charges $e(\boldsymbol{n})$ belonging to the same tetrahedron $T_{\mathscr{M}}$ will occupy the same lattice site exactly once. Since the sum of such complementary charges is zero for neutral cells (see eq 5), these charges cancel each other out. Hence, nonzero complementary charges remain only on the periphery of the nanocluster.

## 3. Details of the Calculations

In the following, we apply the method described in Section 2 to several crystalline lattices. In each case, a lattice unit cell is complemented with a set of charges $e(\boldsymbol{n})$ so as to eliminate several of its multipole moments. These modified unit cells are used to generate a series of finite systems (see Figure 4) as their

$$(2k_1k + 1) \times (2k_2k + 1) \times (2k_3k + 1) \quad k = 0, 1, 2, ... \tag{23}$$

extensions, i.e., a finite system is constructed by repeating the lattice building block $(2k_1k + 1)$, $(2k_2k + 1)$, and $(2k_3k + 1)$ times along the lattice vectors $\boldsymbol{a}_1$, $\boldsymbol{a}_2$, and $\boldsymbol{a}_3$, respectively. Thus, parameters $k_1$, $k_2$, and $k_3$ define the shape of the cluster, and parameter $k$ defines its size.

The convergence of the EP is investigated as a function of the system size and shape, the largest eliminated multipole moment $\mathscr{M}$, the number of tetrahedra $T^{\alpha}_{\mathscr{M}}$, and the values of the shift parameter $v$.

To assess the convergence of the EP, we calculate the potential produced by all centers of the system at the atom sites of the central unit cell, i.e., the *on-site* potential ($V^{\text{site}}$) and the potential in the space *between* the atoms, which is calculated on a three-dimensional grid ($V^{\text{grid}}$) in the central unit cell. In this work, we used a regular grid of $21^3$ points, from which we removed the points if they are within 0.6 Å of any lattice atom. Thus, the total number of remaining grid points was close to 8000.

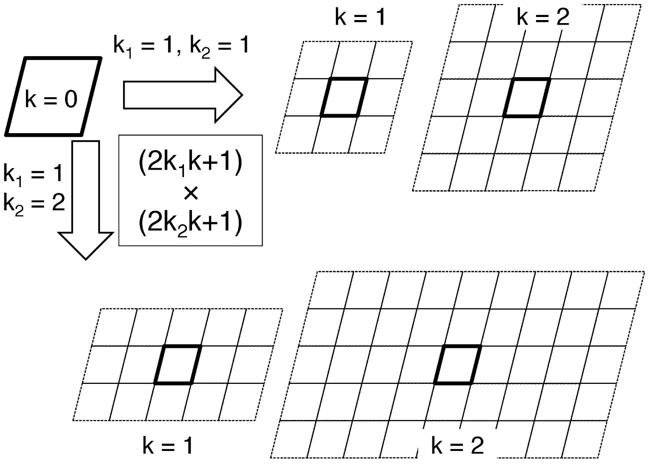


***Figure 4.*** Schematics of the finite clusters constructed according to the $(2k_1k{+}1) \times (2k_2k{+}1)$ rule for $k_1 = k_2 = 1$ (top) and $k_1 = 1$ and $k_2 = 2$ (bottom). The central unit cell is shown with bold lines.

To characterize convergence of the EP, we consider deviations from the reference potential calculated using the Ewald method ($V^{\text{Ewald}}$) for each point $i$

$$\Delta V_i = V_i^{\text{Ewald}} - V_i \tag{24}$$

and the root-mean-square (rms) and the standard deviation of $\Delta V_i$:

$$\Delta V_{\text{rms}} = \sqrt{\frac{1}{N}\sum_{i}^{N}(\Delta V_i)^2} \tag{25}$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i}^{N}(\Delta V_i - \overline{\Delta V})^2} \tag{26}$$

where

$$\overline{\Delta V} = \sum_{i}^{N}\Delta V_i \tag{27}$$

is the mean of $\Delta V_i$, and index $i$ runs through all atoms of the central unit cell in the case of the on-site potential ($V = V^{\text{site}}$) and through all grid points in the case of the potential calculated on the grid ($V = V^{\text{grid}}$).

Structural parameters of the materials selected for this study are summarized in Table 1. The first group includes binary oxides in which the unit cell period is comparable to the interatomic distances: rock-salt MgO, rutile $TiO_2$, and α-quartz $SiO_2$. Due to the high symmetry of the rock-salt lattice, the unit cell used for MgO has zero dipole and quadrupole moments. In the case of rutile, only one component of the dipole moment and three components of the quadrupole moment are equal to zero.

The second group of materials includes complex oxides with ∼50−100 atoms per cell and with a lattice period significantly larger than typical interatomic distances. The EP function in these materials has a complex character combining potential variations on the scale of cation−anion distances with longer range variations on the scale of the lattice period.

## 4. Results

In this section we will investigate the convergence of the EP inside finite systems constructed of building blocks, for which the first $\mathscr{M}$ electric multipole moments are exactly zero. In Section 4.1, we demonstrate that if $\mathscr{M} < 2$, then the EP depends on the shape of the finite systems and illustrate how the absolute convergence of the EP can achieved using larger $\mathscr{M}$. The dependence of the EP on the parameter $v$ (see eq 6) is discussed in Section 4.2. In Section 4.3, we demonstrate that using charges $e(\boldsymbol{n})$ of eight tetrahedra $T^{\alpha}$ can improve the EP convergence. Finally, in Section 4.4,

***Table 1.*** Crystal Lattices Considered in This Work[a]

| | | cell parameters | | | | | |
|---|---|---|---|---|---|---|---|
| material | structure | *a* | *b* | *c* | α | *β* | *γ* |
| MgO | rock-salt | 4.0 | 4.0 | 4.0 | 90.0 | 90.0 | 90.0 |
| $TiO_2$ | rutile | 4.59373 | 4.59373 | 2.95812 | 90.0 | 90.0 | 90.0 |
| $SiO_2$ | α-quartz | 4.91304 | 4.91304 | 5.40463 | 90.0 | 90.0 | 120.0 |
| $Ba_4Fe_8Si_8O_{28}$ | andremeyerite | 7.464 | 13.794 | 7.093 | 90.0 | 118.25 | 90.0 |
| $Na_7[Al_4\ Si_{12}]Si_8O_{48}Cl_3$ | marialite | 12.047 | 12.047 | 7.5602 | 90.0 | 90.0 | 90.0 |
| $Na_{16}Ce_8[CO_3]_{20}$ | petersenite (Ce) | 20.872 | 6.367 | 10.601 | 90.0 | 120.5 | 90.0 |

[a] Structural parameters *a*, *b*, *c* (in Å) and α, *β γ* (in degrees) are given according to the crystallographic convention.

we investigate the EP convergence in the lattices of minerals, which have large unit cells and complex character of the EP function.

**4.1. Size- and Shape-Dependence of The EP.** To illustrate the dependence of the EP inside a finite system on the shape and the size of this system, we applied the $(2k_1k+1) \times (2k_2k+1) \times (2k_3k+1)$ rule, schematically illustrated in Figure 4, to construct a series of rutile $TiO_2$ clusters. We note that for a fixed set of numbers $k_i$ ($k_i \geq 1$, $i = 1, 2, 3$) increasing parameter $k$ to infinity corresponds to the summation over the infinite lattice, while parameters $k_1$, $k_2$, and $k_3$ define a particular order of this summation.

In this set of the calculations, we used eight tetrahedra $T^{\alpha}_{\mathcal{M}}$ and fixed the value of the parameter $v$ (see eq 6) to $v = 0.5$. We considered three sets of $k_1$, $k_2$, and $k_3$:

$$k_1 = 1, \quad k_2 = 1, \quad k_3 = 1$$
$$k_1 = 1, \quad k_2 = 1, \quad k_3 = 2$$
$$k_1 = 1, \quad k_2 = 2, \quad k_3 = 2$$

and varied parameter $k$ from 1 to 5. The dependence of $\Delta V^{\text{site}}_{\text{rms}}$ on $k$ and $k_i$, calculated for several different $\mathcal{M}$, is shown in Figure 5.

We emphasize that the original rutile $TiO_2$ unit cell has two nonzero components of the dipole moment [$Q(0, 1, 0) = Q(0, 0, 1) = -9.2\ e$Å] and three nonzero components of the quadrupole moment [$Q(0, 2, 0) = Q(2, 0, 0) = -32.2\ e\text{Å}^2$, $Q(1, 1, 0) = -16.4\ e\text{Å}^2$]. The procedure described in Section 2 was used to complement the original unit cell with point charges and, thus, to generate rutile $TiO_2$ lattice building blocks, which have no nonzero multipole moments up to $\mathcal{M} = 4$.

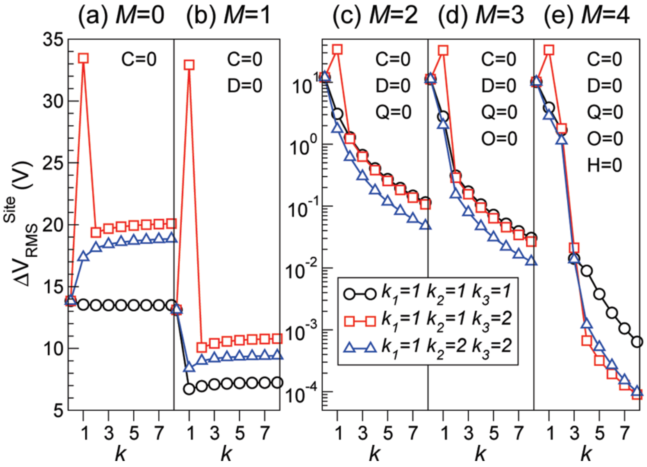


***Figure 5.*** Convergence of the on-site EP calculated for the central unit cell of $(2k_1k + 1) \times (2k_2k + 1) \times (2k_3k + 1)$ rutile $TiO_2$ clusters. $\mathcal{M}$ is the largest eliminated electric moment. Letters *C*, *D*, *Q*, *O*, and *H* refer to the charge, dipole, quadrupole, octopole, and hexadecapole moments of the modified unit cell, respectively.

The results shown in Figure 5a demonstrate that for $\mathcal{M} = 0$, the $\Delta V^{\text{site}}_{\text{rms}}$ converges to a constant value with an increasing value of $k$. The corresponding standard deviation $\sigma^{\text{site}}$ (eq 26) also converges to a constant and nonzero value. For example, for $k = 5$, $\sigma^{\text{site}} = 13.3$ V, if $k_1 = k_2 = k_3 = 1$, $\sigma^{\text{site}} = 19.7$ V, if $k_1 = k_2 = 1$ and $k_3 = 2$, and $\sigma^{\text{site}} = 18.5$ V, if $k_1 = 1$ and $k_2 = k_3 = 2$. In other words, the difference between $V^{\text{site}}$ and $V^{\text{Ewald}}$ varies from site to site and cannot be improved by further increasing $k$. In addition, the convergence limit depends on the choice of $k_i$, i.e., the potential distribution inside the finite cluster is determined by its shape even in the case of an infinitely large $k$. This is characteristic to a series in which the result of the summation depends on the order of this summation, which, in our case, is defined by the parameters $k_i$.

If the original unit cell is complemented with extra charges so as its dipole, but not quadrupole, moment becomes zero ($\mathcal{M} = 1$), the EP still depends on the order of the summation, as demonstrated in Figure 5b. Indeed, series $\sim 1/r^3$ diverges as $\ln(R)$ when summed up over a sphere of radius $R$. In practice, convergence is achieved by fixing the order of the summation. In the case of $TiO_2$ (see Figure 5b), we found that dispersion $\sigma^{\text{site}}$ tends to zero as the value of $k$ is increased, i.e., the EP in the finite system differs from the corresponding $V^{\text{Ewald}}$ by a constant value. We emphasize that the numerical value of that shift is defined entirely by the parameters $k_i$ ($i = 1, 2, 3$) and, in general, by the shape of the finite system.

Once the unit cell is modified so as its quadrupole moment is eliminated ($\mathcal{M} = 2$), the EP converges to $V^{\text{Ewald}}$ with increasing $k$ absolutely, as illustrated in Figure 5c. The speed of the convergence still depends on the parameters $k_i$, i.e., on the shape of the finite system. Similar results, but faster convergence rates, were obtained for $\mathcal{M} = 3$ and 4 (Figure 5d−e).

**4.2. Dependence on the Shift Parameter *v*.** The method described in Section 2 leaves one free to choose the value of the parameter $v$, which defines the shift of the tetrahedra $T^{\alpha}_{\mathcal{M}}$ from their respective corners of the unit cell. Hence, this parameter can be varied in order to minimize the number of the charges and $\Delta V_{\text{rms}}$ for a given value of $\mathcal{M}$ and to improve the convergence of the EP.

The maximum number of charges $e(\mathbf{n})$ associated with each tetrahedron is defined by the value of $\mathcal{M}$ as $n_q = (\mathcal{M} + 1)(\mathcal{M} + 2)(\mathcal{M} + 3)/6$. Hence, the maximum total number

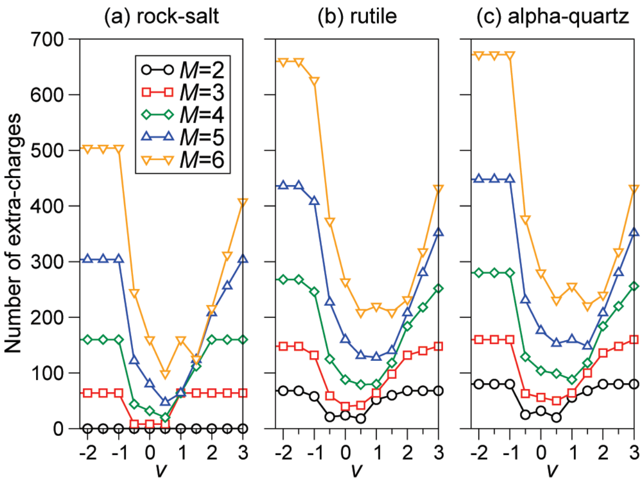


***Figure 6.*** The total number of complementary charges $e(\boldsymbol{n})$ associated with $T^{\alpha}_{\mathscr{M}}$ ($\alpha = 1, ..., 8$) calculated for the unit cells of rock-salt, rutile, and α-quartz lattices and plotted as a function of the shift parameter *v*.

of the charges is given by the number of tetrahedra used multiplied by $n_q$.

However, for all integer and half-integer values of $v$ some of the charges, belonging to different tetrahedra, coincide and may cancel each other out due to the lattice symmetry. For illustration, we plot the total number of charges, $N_q$, for rock-salt, rutile, and α-quartz lattices and several values of $\mathscr{M}$ in Figure 6. Eight tetrahedra were used in each case, which gives the maximum total number of charges $8n_q = 80$ for $\mathscr{M} = 2$, 160 for $\mathscr{M} = 3$, and 280 for $\mathscr{M} = 4$.

As is it clear from Figure 6, the number of charges $N_q$ is considerably smaller than $8n_q$ for the highly symmetrical rock-salt lattice (Figure 6a) but not so for lower symmetry rutile and α-quartz (Figure 6b and c, respectively). This is because many components of the electric multipole moments are equal to zero in the rock-salt structure, which translates into zero values of many of the charges $e(\boldsymbol{n})$.

For $v \geq -0.5$, positions of some of the charges $e^{\alpha}(\boldsymbol{n})$ belonging to different $T^{\alpha}_{\mathscr{M}}$ coincide, and their values can cancel each other out exactly. Consequently, the number of these charges can decrease by as much as a factor of 3 in α-quartz and by a factor of 5 in rock-salt lattices.[49]

The dependence of $\Delta V^{\rm grid}_{\rm rms}$ on the shift parameter $v$ and the size parameter $k$ calculated for rutile $TiO_2$ is shown in Figure 7. Here the value of $v$ was varied from $-3$ to $+3$ with increments of 0.1, and the value of $k$ was varied from 0 to 8.

It is clear that for all $v$ and any $\mathscr{M}$ ($\mathscr{M} \geq 2$), the function $\Delta V^{\rm site}_{\rm rms}$ $(v, k)$ converges to zero in the limit of large $k$. For $\mathscr{M} = 2$ and 3, this function has a narrow deep valley, indicating that the convergence of the EP can be significantly improved by choosing an appropriate $v$. Interestingly, for $\mathscr{M} = 4-7$, we obtain a relatively wide valley, where $\Delta V^{\rm site}_{\rm rms}$ $(v, k)$ is small and almost independent of $v$. This suggests that for large values of $\mathscr{M}$, the EP, due to the modified unit cell, is short range. Hence, the EP inside a finite system converges with its size, as defined by $k$, quickly. At the same time, geometrical size of each $T^{\alpha}_{\mathscr{M}}$ becomes large compared to the size of the original unit cell, which makes the results less dependent on the details of their relative geometrical arrangement, as given by $v$. Interestingly, the range of $v$ providing fast convergence of the EP coincides with that for which $N_q$ is the smallest.

To summarize, parameter $v$ determines the shift of the extra charges from the unit cell corners and serves two purposes. First, the total number of extra charges can be reduced significantly if $v > 0$, as illustrated in Figure 6. Second, positive $v$ can improve convergence of the electrostatic with respect to the highest eliminated multipole moment and the size of the nanocluster (Figure 7). In addition, parameter $v$ offers flexibility in positioning the extra charges, with respect to the atomic coordinates, which may be of an advantage in modeling surface sites.

**4.3. Dependence on Spatial Distribution of the Charges $\boldsymbol{e(n)}$.** The charges $e^{\alpha}(\boldsymbol{n})$ associated with *any single* tetrahedron $T^{\alpha}_{\mathscr{M}}$ are sufficient to eliminate all electric moments of the unit cell up to any predefined $\mathscr{M}$.[36]

In this section, we investigate the effects of symmetrical spatial distribution of the charges $e(\boldsymbol{n})$ generated for several tetrahedra. For example, it can be expected that the charges associated with eight tetrahedra positioned symmetrically at the corners of the unit cell, as indicated in Figure 8c, could provide faster convergence of the EP than those due to a single tetrahedron (see Figure 8a).

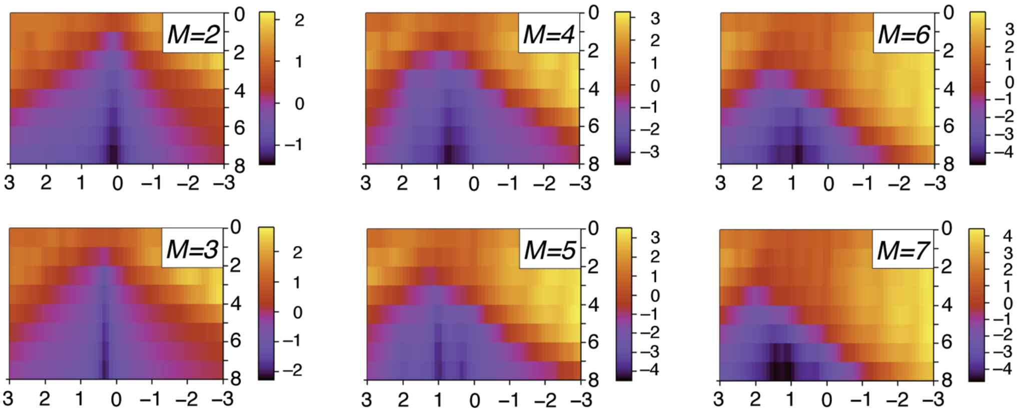


***Figure 7.*** Convergence of $\Delta V^{\rm grid}_{\rm rms}$ in rutile $TiO_2$ calculated for several values of $\mathscr{M}$ and the shift parameter *v*: $-3 \leq v \leq 3$. Darker regions indicate smaller values of $\Delta V^{\rm grid}_{\rm rms}$; the corresponding range of values is shown on the right of each plot using a $\log_{10}$ scale.

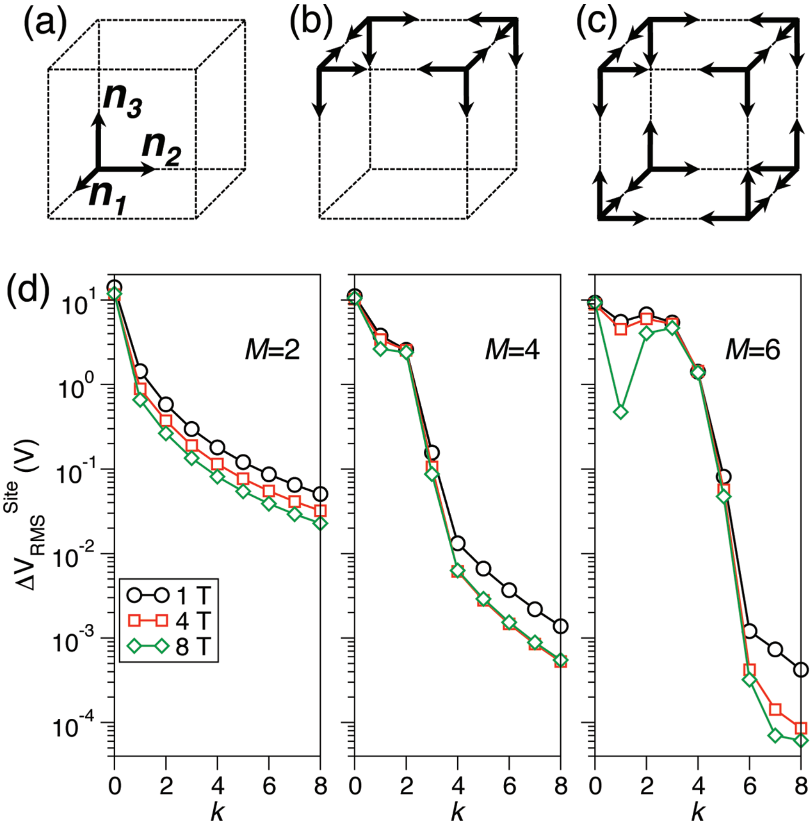


*Figure 8.* Convergence of the on-site EP in α-quartz. The values of $\Delta V_{\mathrm{rms}}^{\mathrm{site}}$ are calculated for clusters generated using eq 23 with $k_1 = k_2 = k_3 = 1$. The charges are generated using one (1 T), four (4 T) and eight (8 T) tetrahedra as indicated in a, b, and c, respectively. The orientation of each tetrahedron is given by vectors $\boldsymbol{n}_1$, $\boldsymbol{n}_2$, and $\boldsymbol{n}_3$, which are collinear with the lattice vectors $\boldsymbol{a}_1$, $\boldsymbol{a}_2$, and $\boldsymbol{a}_3$ (not shown) of the unit cell, indicated with a dashed box in a−c. $\mathscr{M}$ shows the value of the largest eliminated multipole.

We consider the α-quartz $SiO_2$ lattice and construct the charges $e(\boldsymbol{n})$ using one, four, and eight tetrahedra as indicated in Figure 8, in order to eliminate the electric moments up to $\mathscr{M} = 2$, 4, and 6 in each case. A series of finite clusters was constructed using the $(2k + 1)^3$ rule, given by eq 23 with $k_1 = k_2 = k_3 = 1$, for $k = 0,1, ..., 8$, and the values of $\Delta V_{\mathrm{rms}}^{\mathrm{site}}$ were calculated for the central cell of each cluster. The results of these calculations for the shift vector $v = 0.0$ are plotted in Figure 8d.

It is clear that even in the case of the low-symmetry lattice, such as that of α-quartz, the convergence of the EP can be improved if several $T_{\mathscr{M}}^{\alpha}$ tetrahedra of charges are used. We note that, in this particular case, the $e(\boldsymbol{n})$ constructed for the four and eight tetrahedra demonstrate almost identical behavior of $\Delta V_{\mathrm{rms}}^{\mathrm{site}}$ with the value of $k$.

The advantages of using eight symmetrically located tetrahedra becomes apparent if we consider the dependence of $\Delta V_{\mathrm{rms}}^{\mathrm{site}}$ on the shift vector $v$, as shown in Figure 9. Here we plot $\Delta V_{\mathrm{rms}}^{\mathrm{site}}$ as a function of $v$ for three sizes of the α-quartz clusters ($k = 4, 6, 8$) and $\mathscr{M} = 4$ and 6.

In the case of a single tetrahedron, $\Delta V_{\mathrm{rms}}^{\mathrm{site}}$ strongly depends on $v$, i.e., achieving good EP convergence requires preoptimization of the shift parameter. The dependence on the value of $v$ is less pronounced in the case of four tetrahedra. Finally, if eight tetrahedra are used, the $\Delta V_{\mathrm{rms}}^{\mathrm{site}}$ shows small variations near its minimum, which is broadly in the region of $0 \leq v \leq 1$.

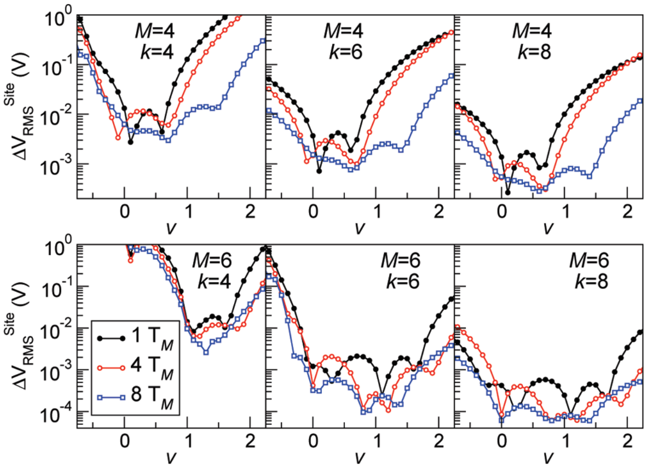


*Figure 9.* Dependence of $\Delta V_{\mathrm{rms}}^{\mathrm{site}}$ in the central cell of α-quartz clusters on the shift parameter $v$. Complementary charges $e(\boldsymbol{n})$ are generated using one (closed circles), four (open circles), and eight (open squares) tetrahedra $T_{\mathscr{M}}$ for $\mathscr{M} = 4$ (top panels) and 6 (bottom panels). The clusters are generated using eq 23 with $k_1 = k_2 = k_3 = 1$.

This observation is consistent with the results obtained for $\Delta V_{\mathrm{rms}}^{\mathrm{grid}}$ in rutile $TiO_2$ (Section 4.2) and suggests that the particular choice of $v$ is insignificant for $\mathscr{M} \geq 4$ and $k \geq 4$ as long as $0 \leq v \leq 1$.

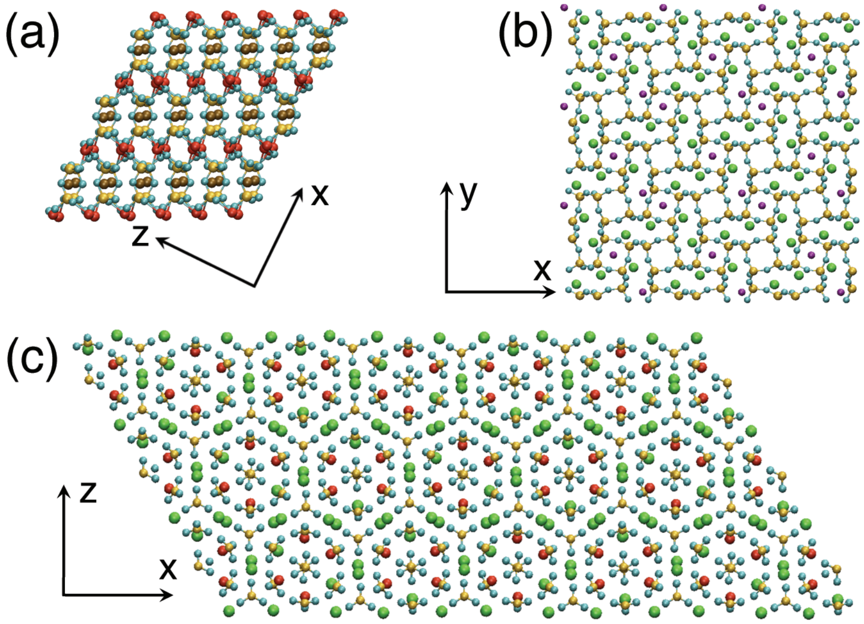


***Figure 10.*** Structure of complex oxides used in this work: (a) andremeyerite $Ba_4Fe_8Si_8O_{28}$; (b) marialite $Na_7[Al_4Si_{12}]Si_8O_{48}Cl_3$, and (c) petersenite $Na_{16}Ce_8[CO_3]_{20}$. Each panel shows a $(2k + 1)^3$ extension of the corresponding unit cell with $k = 1$. Structural parameters of these systems are given in Table 1.

**4.4. Electrostatic Potential in Complex Lattices.** To illustrate the applicability of the method to a wide range of systems, we considered three minerals having complex lattice structures formed by several types of atoms with different formal ionic charges (see Table 1 and Figure 10).

In particular, andremeyerite has a monoclinic lattice and 48 ions per cell with formal charges of −2, +2, and +4. Marialite has tetragonal lattice, and its unit cell contains 82 ions with the charges of −2, −1, +1, +3, and +4. Finally, petersenite has a monoclinic lattice and contains 20 molecular anions $CO_3$ per cell. These anions were modeled using the corresponding formal charges as $C^{4+}O_3^{2-}$ anions, which provide strong variations of the EP on the scale of the interatomic distances.

The EP distribution in these systems is further complicated by the long-range modulations on the scale of the lattice period and by large differences in the values of the crystallographic parameters. For example, in petersenite, the value of the lattice parameter *a* is ~20.872 Å and the ratios *a*/*b* and *a*/*c* are ~3.2 and ~2.0, respectively.

In each case, we have modified the initial unit cell in order to eliminate $\mathscr{M}$ lowest multipole moments and used these modified cells to generate finite $(2k + 1)^3$ systems, as described above. Convergence of the EP was investigated as a function of the system size. In all cases, we used eight $T^{\alpha}_{\mathscr{M}}$ tetrahedra, and the value of the shift parameter $v$ was fixed to zero. The results of these calculations are collected in Figure 11.

The calculated function $\Delta V^{\text{site}}_{\text{rms}}$ $(k, \mathscr{M})$ does not generally approach zero if the cell is neutral and has no dipole moment, simultaneously, i.e., $\mathscr{M} = 1$. Instead, $\Delta V^{\text{site}}_{\text{rms}}$ converges to a constant value, as it is shown in Figure 11a−c, which depends on the order of summation, as discussed above in Section 4.1 on the example of rutile $TiO_2$. The convergence becomes absolute if the quadrupole moment is eliminated as well, i.e., $\mathscr{M} \geq 2$.

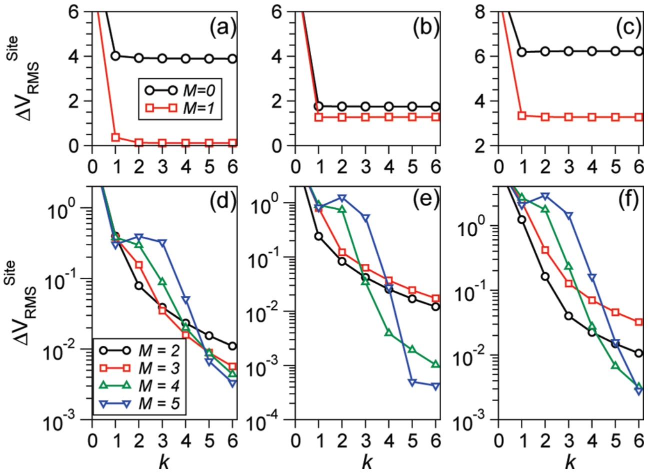


***Figure 11.*** Convergence of the on-site EP in andremeyerite (a, d), marialite (b, e), and petersenite (c, f) clusters generated using eq 23 with $k_1 = k_2 = k_3 = 1$. The original unit cells have been modified to eliminate all electric moments up to and including $\mathscr{M}$. In the case of unit cells with nonzero quadrupole ($\mathscr{M} < 2$), the potential converges to an arbitrary limit defined by the shape of the cluster (a−c). The absolute convergence is achieved if $\mathscr{M} \geq 2$ (d−e).

The convergence with respect to the system size is faster for larger values of $\mathscr{M}$, as illustrated in Figure 11d−f. We notice that, in some cases, functions $\Delta V^{\text{site}}_{\text{rms}}$ $(k, \mathscr{M})$ calculated for an even $\mathscr{M} = \mathscr{M}_1$ and odd $\mathscr{M} = \mathscr{M}_1 + 1$ behave similarly with *k*. This can be seen for both marialite and petersenite for $\mathscr{M} = 2$ and 3 and for $\mathscr{M} = 4$ and 5. This is because in the case of even $\mathscr{M}$ some components of the higher electric moments can become eliminated by symmetry, and hence, further increase of $\mathscr{M}$ by 1 may improve the convergence insignificantly.

The overall convergence of the EP in the considered minerals is similar to that found for simpler structures, such as rutile and α-quartz. Similarly good convergence was found for the EP calculated on the grid and characterized using $\Delta V_{\rm rms}^{\rm grid}$ $(k, \mathscr{M})$.

## 5. Discussion and Conclusions

The relation between the electrostatic potential (EP) inside a finite macroscopic sample and its surface charge as well as the dependence of the potential on the shape of the sample have been considered previously in, for example, refs 50−52. It is well-known that surfaces of macroscopic samples acquire surface charge in order to compensate the EP *outside* the sample. There can be several sources of the surface charge, including surface reconstruction, contamination with impurities, and defect formation, all of which can be difficult to describe on the atomic scale.

In our method, the lattice unit cell is complemented with charges $e(\boldsymbol{n})$ so as the EP produced by each modified cell is short range, which is the physical basis for the proposed regularization of eq 1. Indeed, the EP *outside* a finite system constructed from these unit cells is also short range independently on its size and shape, as expected for realistic macroscopic samples. It can also be said that charges $e(\boldsymbol{n})$ produce effective compensating potential, which does not need to be described on the atomic scale.

The potential *inside* a finite system converges with its size absolutely, and the rate of convergence is controlled by few numerical parameters: the largest eliminated multipole moment of the original unit cell $\mathscr{M}$, the number of the tetrahedra $T_{\mathscr{M}}$ of the complementary charges, and the displacement $v$ of the main tetrahedra vertices with respect to the unit cell corners. We note that the absolute convergence of the EP with the size of the system is achieved only if $\mathscr{M} \geq 2$ and can be improved by increasing $\mathscr{M}$ further, by using a symmetry adjusted number of tetrahedra and by varying the value of $v$. As demonstrated above, elimination of the unit cell dipole moment only ($\mathscr{M} = 1$) does not eliminate the dependence of the EP on the shape of the finite system. Importantly, as the size of the finite system increases, the EP inside of it converges to the result of the Ewald summation for the corresponding infinite lattice.[36] Thus, regularization of the Coulomb series proposed in our method is equivalent to that used in the Ewald method.

We note that nonzero complementary charges $e(\boldsymbol{n})$ are situated only near the periphery of the nanocluster, while in its inner region, the lattice remains unchanged. This property of $e(\boldsymbol{n})$, together with the rapid convergence of the EP with the size of the nanocluster, makes this method convenient to use in embedded cluster calculations of crystalline materials. This method can be also used for disordered materials and liquids as well as low-dimensional systems (surfaces, quasi-one-dimensional wires, and clusters), providing they can be represented using a supercell approach. Indeed, a supercell can be considered on the same footing as a conventional crystalline cell, and the same formalism can be applied. We can add that the supercell, complemented with the extra-charges $e(\boldsymbol{n})$, can be considered as a lattice building block and can be used to construct arbitrary finite structures consistent with the problem at hand.

The computational cost of generation charges $e(\boldsymbol{n})$ for a unit cell depends on the largest multipole moment to be eliminated and on the number of particles in the unit cell and scales as $\mathscr{M}^3\mathscr{N}_0$. The unit cell complemented with the extra charges is used to construct a nanocluster of $(2k + 1)^3$ unit cells, which makes the total number of atoms in the system $N = (2k + 1)^3\mathscr{N}_0$; the number of nonzero extra charges at the surface of the nanoclusters scales as $\mathscr{M}^3k^2$. Thus, the dominant contribution to the cost of the calculating the EP at a single point scales linearly with the number of atoms in the nanocluster.

To summarize, we suggest a systematic way of constructing accurate electrostatic embedding potential for bulk, surfaces, and nanostructures of crystalline materials. To regularize the summation of the EP series, the original lattice cell is modified so as to eliminate its electric multipole moments up to any given $\mathscr{M}$ and, thus, to make the potential produced by such a cell short range. We applied this method to several crystals, including complex minerals containing 50−100 atoms per cell, and demonstrated rapid convergence of the EP inside finite nanoscale clusters. The method is fully analytical, and the convergence can be controlled by a few well-defined numerical parameters.

**Acknowledgment.** The authors are grateful to A. L. Shluger for stimulating discussions and comments on the manuscript. P.V.S. is supported by the Royal Society. I.V.A.'s stay at University College London was supported by the Thomas Young Centre—the London Centre for Theory and Simulation of Materials—and by the EPSRC Grant no. GR/S80080/01. The calculations have been performed on the computer cluster funded via the Elements Science and Technology Project, MEXT, and Grant-in-Aid for Young Scientists 20810004 at the WPI-Advanced Institute for Materials Research, Tohoku University, Japan.

### References

(1) Stefanovich, E. V.; Shidlovskaya, E. K.; Shluger, A. L.; Zakharov, M. A. *Phys. Stat. Sol. B* **1990**, *160*, 529.

(2) Abarenkov, I. V.; Bulatov, V. L.; Godby, R.; Heine, V.; Payne, M. C.; Souchko, P. V.; Titov, A. V.; Tupitsyn, I. I. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1997**, *56*, 1743.

(3) Govind, N.; Wang, Y. A.; Carter, E. A. *J. Chem. Phys.* **1999**, *110*, 7677.

(4) Donnerberg, H.; Birkholz, A. *J. Phys.: Condens. Matter* **2000**, *12*, 8239.

(5) Sushko, P. V.; Shluger, A. L.; Catlow, C. R. A. *Surf. Sci.* **2000**, *450*, 153.

(6) Sherwood, P.; de Vries, A. H.; Guest, M. F.; Schreckenbach, G.; Catlow, C. R. A.; French, S. A.; Sokol, A. A.; Bromley, S. T.; Thiel, W.; Turner, A. J.; Billeter, S.; Terstegen, F.; Thiel, S.; Kendrick, J.; Rogers, S. C.; Casci, J.; Watson, M.; King, F.; Karlsen, E.; Sjovoll, M.; Fahmi, A.; Schfer, A.; Lennartz, C. *J. Mol. Struct.* **2003**, *632*, 1.

(7) Nasluzov, V. A.; Ivanova, E. A.; Shor, A. M.; Vayssilov, G. N.; Birkenheuer, U.; Rösch, N. *J. Phys. Chem. B* **2003**, *107*, 2228.

(8) Herschend, B.; Baudin, M.; Hermansson, K. *J. Chem. Phys.* **2004**, *120*, 4939.

(9) Seijo, L.; Barandiarán, Z. *J. Chem. Phys.* **2004**, *121*, 6698.

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

(11) Danyliv, O.; Kantorovich, L.; Cora, F. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2007**, *76*, 045107.

(12) Kästner, J.; Thiel, S.; Senn, H. M.; Sherwood, P.; Thiel, W. *J. Chem. Theory Comput.* **2007**, *3*, 1064.

(13) Higashi, M.; Truhlar, D. G. *J. Chem. Theor. Comput.* **2008**, *4*, 790.

(14) Burow, A. M.; Sierka, M.; Döbler, J.; Sauer, J. *J. Chem. Phys.* **2009**, *130*, 174710.

(15) Kamerlin, S. C. L.; Haranczyk, M.; Warshel, A. *J. Phys. Chem. B* **2009**, *113*, 1253.

(16) Lin, H.; Truhlar, D. G. *Theor. Chem. Acc.* **2007**, *117*, 185.

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

(18) McKenna, K. P.; Sushko, P. V.; Shluger, A. L. *J. Am. Chem. Soc.* **2007**, *129*, 8600.

(19) Kimmel, A. V.; Sushko, P. V.; Shluger, A. L.; Kuklja, M. M. *J. Phys. Chem. A* **2008**, *112*, 4496.

(20) Torras, J.; Bromley, S.; Bertran, O.; Illas, F. *Chem. Phys. Lett.* **2008**, *457*, 154.

(21) Shluger, A. L.; Sushko, P. V.; Kantorovich, L. N. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1999**, *59*, 2417.

(22) Govind, N.; Sushko, P. V.; Hess, W. P.; Valiev, M.; Kowalski, K. *Chem. Phys. Lett.* **2009**, *470*, 353.

(23) Pacchioni, G.; Valentin, C. D.; Dominguez-Ariza, D.; Illas, F.; Bredow, T.; Kluner, T.; Staemmler, V. *J. Phys.: Condens. Matter* **2004**, *16*, S2497.

(24) Sousa, C.; de Graaf, C.; Lopez, N.; Harrison, N. M.; Illas, F. *J. Phys.: Condens. Matter* **2004**, *16*, S2557.

(25) Giordano, L.; Sushko, P. V.; Pacchioni, G.; Shluger, A. L. *Phys. Rev. Lett.* **2007**, *99*, 136801.

(26) Müller, M.; Stankic, S.; Diwald, O.; Knözinger, E.; Sushko, P. V.; Trevisanutto, P. E.; Shluger, A. L. *J. Am. Chem. Soc.* **2007**, *129*, 12491.

(27) Bo, C.; Maseras, F. *Dalton Trans.* **2008**, *22*, 2911.

(28) Kantorovich, L. N.; Shluger, A. L.; Sushko, P. V.; Günster, J.; Stracke, P.; Goodman, D. W.; Kempter, V. *Faraday Discuss.* **1999**, *114*, 173.

(29) Sushko, P. V.; Gavartin, J. L.; Shluger, A. L. *J. Phys. Chem. B* **2002**, *106*, 2269.

(30) Ricci, D.; Valentin, C. D.; Pacchioni, G.; Sushko, P. V.; Shluger, A. L.; Giamello, E. *J. Am. Chem. Soc.* **2003**, *125*, 738.

(31) Sterrer, M.; Diwald, O.; Knözinger, E.; Sushko, P. V.; Shluger, A. L. *J. Phys. Chem. B* **2002**, *106*, 12478.

(32) Sterrer, M.; Berger, T.; Diwald, O.; Knözinger, E.; Sushko, P. V.; Shluger, A. L. *J. Chem. Phys.* **2005**, *123*, 064714.

(33) Beck, K. M.; Henyk, M.; Wang, C.; Trevisanutto, P. E.; Sushko, P. V.; Hess, W. P.; Shluger, A. L. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2006**, *74*, 045404.

(34) Beck, K. M.; Joly, A. G.; Diwald, O.; Stankic, S.; Trevisanutto, P. E.; Sushko, P. V.; Shluger, A. L.; Hess, W. P. *Surf. Sci.* **2008**, *602*, 1968.

(35) Contact authors for a stand alone Fortran 77 code generating charges $e(\boldsymbol{n})$ for an arbitrary lattice.

(36) Abarenkov, I. V. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2007**, *76*, 165127.

(37) Evjen, H. M. *Phys. Rev.* **1932**, *39*, 675.

(38) Wolf, D.; Keblinski, P.; Phillpot, S. R.; Eggebrecht, J. *J. Chem. Phys.* **1999**, *110*, 8254.

(39) Sherwood, P.; de Vries, A. H.; Collins, S. J.; Greatbanks, S. P.; Burton, N. A.; Vincent, M. A.; Hillier, I. H. *Faraday Discuss.* **1997**, *106*, 79.

(40) Stefanovich, E. V.; Truong, T. N. *J. Phys. Chem. B* **1998**, *102*, 3018.

(41) French, S. A.; Sokol, A. A.; Bromley, S. T.; Catlow, C. R. A.; Rogers, S. C.; King, F.; Sherwood, P. *Angew. Chem., Int. Ed. Engl.* **2001**, *40*, 4437.

(42) Teunissen, E. H.; Jansen, A. P. J.; van Santen, R. A.; Orlando, F. L.; Dovesi, R. *J. Chem. Phys.* **1994**, *101*, 5865.

(43) Nam, K.; Gao, J.; York, D. M. *J. Chem. Theory Comput.* **2005**, *1*, 2.

(44) Ewald, P. P. *Ann. Phys. (Leipzig)* **1921**, *64*, 253.

(45) Porto, M. *J. Phys. A: Math. Gen.* **2000**, *33*, 6211.

(46) Madelung, E. *Phys. Z.* **1918**, *19*, 524.

(47) Kudin, K. N.; Scuseria, G. E. *Chem. Phys. Lett.* **1998**, *283*, 61.

(48) Ramo, D. M.; Sushko, P. V.; Gavartin, J. L.; Shluger, A. L. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2008**, *78*, 235432.

(49) When the original unit cell is shifted to place the main vertex of a tetrahedron $T^{\alpha}_{\mathcal{M}}$ to the origin, some components of its multipole moments can acidentally become equal to zero. In this case, the number of charges $e(\boldsymbol{n})$ can also reduce. In our calculations, this was the case for the rutile lattice and $v = -1.0$ (see Figure 6b).

(50) Tupizin, I. I.; Abarenkov, I. V. *Phys. Stat. Sol. B* **1977**, *82*, 99.

(51) Smith, E. R. *Proc. R. Soc. London, Ser. A* **1981**, *375*, 475.

(52) Kantorovich, L. N.; Tupitsyn, I. I. *J. Phys.: Condens. Matter* **1999**, *11*, 6159.

CT900480P

JCTC Journal of Chemical Theory and Computation

# Design of a Versatile Force Field for the Large-Scale Molecular Simulation of Solid and Liquid OMCTS

Hiroki Matsubara,*[,†] Fabio Pichierri,*[,†] and Kazue Kurihara[‡]

*G-COE Laboratory, Department of Applied Chemistry, Graduate School of Engineering, Tohoku University and JST-CREST, Aoba-yama 6-6-07, Sendai 980-8579, Japan, and Institute of Multidisciplinary Research for Advanced Materials (IMRAM), Tohoku University, JST-CREST, 2-1-1 Katahira, Sendai 980-8577, Japan*

Received November 17, 2009

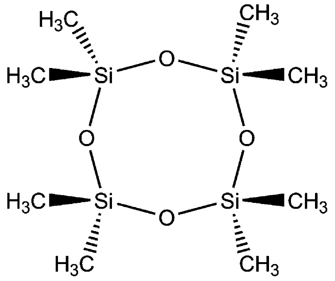
**Abstract:** We developed a new, versatile force field for the molecular simulation of octamethylcyclotetrasiloxane (OMCTS) both in the solid and liquid phases. From a series of molecular dynamics simulations, we obtain good agreement with the experimental lattice constants, sublimation enthalphy, and molecular packing of the crystal. The experimental density, diffusion coefficient, and shear viscosity of this van der Waals liquid in the range 300−440 K are well reproduced as well. The new force field can be thus employed in the large-scale molecular simulation of liquid OMCTS where structural details are important in determining the collective properties of the system.

## I. Introduction

Octamethylcyclotetrasiloxane (OMCTS, Chart 1) is a small macrocycle made of four covalently linked $Si(Me)_2O$ units which has been used as a model liquid in surface force measurements (SFM).[1–7] These experiments have revealed the existence of oscillatory solvation forces which are characteristic of the behavior of liquids confined in nanosized spaces, and OMCTS, due to its quasi-spherical shape and zero dipole moment, has contributed to the development of theoretical treatments of these phenomena. In many cases, the molecular level origin of such phenomena is not clear from these experiments, and molecular simulation, particularly molecular dynamics (MD) and Monte Carlo methods, is used to connect these phenomena with the behavior of molecular ensembles that are subject to confinement.[8–10]

Molecular simulation of real materials requires realistic potentials which are necessary to accurately describe the interactions among a large number of molecules.[11,12] Especially challenging is the design of intermolecular potentials concerned with weak dispersion interactions that operate among large molecules.[13] Flexible molecules further increase the complexity of the potential thereby making large-scale computations very demanding.

***Chart 1***



Possible alternative approaches that overcome system size are multiscale and coarse grain models.[14,15] The former are concerned with the integrated combination of different methodologies, from quantum mechanics up to the finite element method, each of which is appropriate to describe the system under study at different scales. Coarse grain (CG) methods on the other hand are used to simulate ensembles of large molecules each of which is modeled as a collection of beads rather than atoms. In this regard, Klein and coworkers have recently performed CG-MD simulations to investigate antimicrobial polymers and polypeptides.[16] Another difficulty encountered in the large-scale computations is concerned with the transferability of the interatomic or

* Corresponding author e-mail: fabio@che.tohoku.ac.jp (F.P.); matsubara@che.tohoku.ac.jp (H.M.).

[†] Graduate School of Engineering.

[‡] IMRAM.

10.1021/ct9006053
Published on Web 03/01/2010

intermolecular potentials to different physical states of the system, namely condensed (solid and liquid) and vapor phases. Therefore, with a few exceptions,[17,18] the large majority of force fields are generally employed to study complex molecular systems that are in a specific phase while the study of different phases of the system usually requires using different sets of parameters.

So far, only simple spherical (or ellipsoidal) models are found in molecular simulation studies of OMCTS.[8–10] For instance, Ayappa and Mishra[10] have recently performed a series of grand canonical Monte Carlo simulations on spherical OMCTS molecules of diameter 7.7 Å and interacting with each other through a 12−6 Lennard-Jones (LJ) potential. While such models are useful for discussing phase transitions that occur under confinement, more desirable is the model that makes use of information contained in the molecular structure so that atomic-scale details on both molecular packing and specific intermolecular interactions can be gained from the simulations. To develop such a model is the main objective of this study.

The paper is structured as follows. In section II, we develop our original force field based on the assumption that the methyl−methyl (Me−Me) interaction is the dominant intermolecular interaction among OMCTS molecules. *Ab-initio* quantum mechanical (QM) calculations are employed to construct the potential energy surface corresponding to the interaction between methane molecules ($CH_4$). This intermolecular potential is expected to mimic the weak van der Waals (vdW) interactions that operate among the methyl groups of OMCTS molecules. We subsequently perform a series of classical MD simulations[11,12] to test the new potential and hence confirm our initial assumption. The potential is validated in section III by checking how accurately the crystal lattice constants, sublimation enthalpy ($\Delta H_{\mathrm{subl}}$), and liquid density (at 300 and 400 K) are reproduced. In section IV, the force field is further refined by employing a penalty function which depends on the lattice constants, $\Delta H_{\mathrm{subl}}$, and the density of the liquid at two different temperatures. In section V, the refined force field is then employed in a long MD simulation of OMCTS liquid, and the calculated bulk properties are compared against experimental data. Conclusions and remarks are given in section VI.

## II. Potential Model

The OMCTS molecule has a disk-like shape and possesses eight methyl groups which are located at the outermost point while the atoms of the siloxane ring (Si and O) are well embedded inside the van der Waals (vdW) surface of the molecule, as shown in Figure 1. Besides shielding the siloxane ring from the outer environment, the methyl groups also confer overall conformational rigidity to the macrocyle owing to the increase in intramolecular steric repulsion. Because OMCTS is a neutral molecule with a negligible dipole moment ($\mu = 0.22$ D), dispersive interactions are likely the main components of the intermolecular interactions among OMCTS molecules. An inspection into the OMCTS crystal[19] indicates that the shortest intermolecular distance among non-hydrogen atoms is that between the carbon atoms of the methyl groups (C···C < 4.0 Å) while CH···HC contacts are in the range 2.3−3.0 Å, thereby supporting the above hypothesis. We therefore considered a model of the OMCTS molecule in which only Me−Me interactions are included. The molecule is assumed to be a rigid body whose geometry is taken from the molecular crystal.[19] One methyl group site is located on each carbon atom position, and hydrogen atoms were implicitly included. The interaction energy $E_{ij}$ of a pair of methyl sites $i$ and $j$ in different molecules is described by the Lennard-Jones potential:

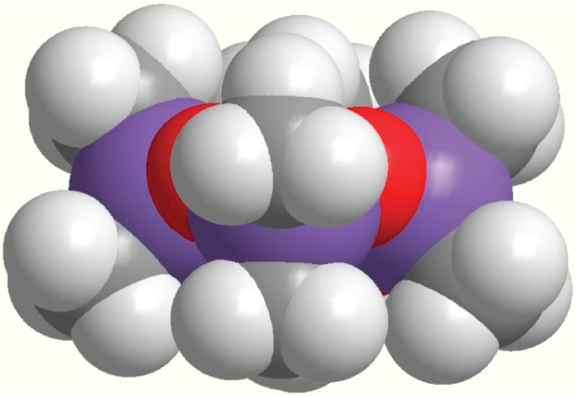


*Figure 1.* Space-filling representation of the OMCTS molecule.

$$E_{ij} = 4\varepsilon\left[\left(\frac{\sigma}{r_{ij}}\right)^{12} - \left(\frac{\sigma}{r_{ij}}\right)^{6}\right] \tag{1}$$

where, $r_{ij}$ is the distance between methyl sites. The parameters $\sigma$ and $\varepsilon$ were derived from a series of QM calculations performed on the methane dimer.

The quantum mechanical approach employed here is based on the second-order Møller−Plesset perturbation theory (MP2)[20] in combination with the 6-31G(d,p),[21] 6-311++G(2d,2p),[22] and AUG-cc-pVDZ[23] basis sets as implemented in the parallel version of the Gaussian 03 software package.[24] Early theoretical studies on the methane dimer have established that the lowest energy configuration arises from the face-to-face interaction between methane molecules in the $D_{3d}$-symmetric dimer.[25] By using this configuration, we first optimized the geometry of the dimer at the three levels of MP2 theory described above. The carbon−carbon distance in the optimized geometries of the dimer correspond to 3.804 Å at the MP2/6-31G(d,p) level, 3.803 Å at the MP2/6-311++G(2d,2p) level, and 3.575 Å at the MP2/AUG-cc-pVDZ level. Starting from these optimized geometries, we have computed the corresponding potential energy curves (PECs) by stepwise elongation and compression of the carbon−carbon distance while optimizing all the remaining degrees of freedom. From these calculations, we notice that only the PEC computed with the smaller 6-31G(d,p) basis set appears as a continuous curve, whereas those computed with the larger basis sets show a discontinuity in the repulsive part. This problem arises from the well-known basis set superposition error (BSSE)[26] which increases dramatically at $r(\mathrm{CC}) < r(\mathrm{CC})_{\mathrm{eq}}$. We have therefore computed the PECs by performing relaxed potential energy surface scans which were corrected at each step by using the counterpoise correction (CP) method of Boys and Bernardi.[27]

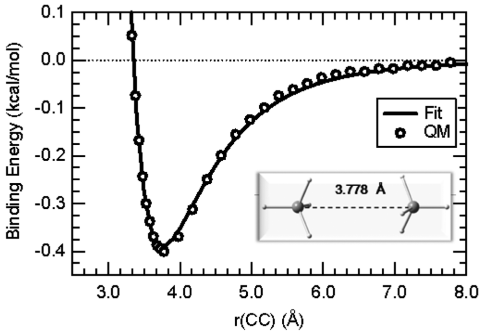


*Figure 2.* CP-corrected potential energy curve for the $D_{3d}$-symmetric methane dimer as computed at the MP2/AUG-cc-pVDZ level of theory. The inset shows the CP-corrected optimized geometry of the methane dimer.

The CP-corrected PEC of the methane dimer computed at the MP2/AUG-cc-pVDZ level of theory is shown in Figure 2. The minimum is located at $r(\mathrm{CC})_{\mathrm{eq}} = 3.778$ Å, and the BSSE-corrected binding energy of the dimer corresponds to 0.402 kcal/mol. By fitting this PEC with the LJ function (eq 1), the parameters $\sigma = 3.35$ Å and $\varepsilon = 0.390$ kcal/mol were derived. This set of parameters is hereafter labeled as Model-A. In comparison, the CP-corrected geometry of the minimum computed at the MP2/6-311++G(2d,2p) level is located at $r(\mathrm{CC})_{\mathrm{eq}} = 3.804$ Å. Also, we checked how the introduction of silicon affects $r(\mathrm{CC})_{\mathrm{eq}}$ by considering the $H_3C-SiH_3$ molecule. The CP-corrected geometry of the $H_3Si-CH_3\cdots H_3C-SiH_3$ dimer obtained at the MP2/AUG-cc-pVDZ level of theory is characterized by $r(\mathrm{CC})_{\mathrm{eq}} = 3.867$ Å, which is only 0.089 Å longer than the corresponding intermolecular carbon−carbon distance of the methane dimer. This small elongation of $r(\mathrm{CC})_{\mathrm{eq}}$ can be attributed to the presence of repulsive dipole−dipole interactions ($\mu = 0.69$ D) that are operative in the $H_3C-SiH_3$ dimer with respect to the methane dimer where $\mu = 0$ D. Despite its simplicity, the methane dimer represents a good starting model for modeling the methyl−methyl interaction between two OMCTS molecules whose dipole moment is only 0.22 D (for the $C_s$-symmetric geometry optimized at the PBE1PBE/6-31G(d) level of theory).

## III. Model Validation

The LJ parameters of Model-A were validated by performing MD simulations on both liquid and crystal phases. We utilized the DLPOLY2 (version 2.20) package[28] for all the MD simulations in this study. For the crystal phase, we executed simulations with a constant number of molecules, constant pressure, and constant temperature where the size and shape of the simulation box were allowed to change ($N\sigma T$ ensemble) by using the method of Berendsen et al.[29] The time constants used were 0.2 ps for the thermostat and 0.3 ps for the barostat. The simulation box contained 216 OMCTS molecules, and three-dimensional periodic boundary conditions (PBCs) were imposed. The cutoff radius for the vdw interactions was 20.0 Å. The time step was set to 2.0 fs. The velocity Verlet and NOSQUISH[30] algorithms were used for the numerical integration of the translational and rotational parts of the equation of motion. The temperature and total pressure were controlled to 223 K and 1 atm. The initial configuration was constructed by replicating the unit cell of the molecular crystal[19] so as to obtain a box of dimensions $A = 3a$, $B = 3b$, $C = 6c$. After a 40 ps run for equilibration, a statistical average was taken over 100 ps.

For the liquid phase, we executed simulations with a constant number of molecules, constant pressure, and constant temperature where only the cell size was allowed to vary (*NPT* ensemble) by using the method of Berendsen et al.[29] The time constants were the same as those of the crystal simulation. We considered the two different temperatures of 300 and 400 K which are close to the experimental melting and boiling temperatures of 290 and 444 K,[31] respectively. For all these cases, the pressure was set to 1 atm. The initial configuration was obtained from a random molecular configuration kept at a temperature of 1000 K for several tens of picoseconds while the simulation box was a cube of 48.3 Å. Subsequently, the system was quenched to the setting temperature of 300 or 400 K by rescaling the molecular velocities, and from this point on the box size was allowed to vary. The statistical average was taken over 200 ps after several tens of picoseconds for equilibration. Other conditions were the same as those employed in the simulation of the crystal. From these simulations, we calculated the lattice constants, *a*, *b*, and *c*; sublimation enthalpy, $\Delta H_{\mathrm{subl}}$ (calculated here as the potential energy per molecule at 223 K); and liquid densities at 300 and 400 K, $\rho_{300}$ and $\rho_{400}$. Those are compared with the experimental values reported in Table 1.

As can be noted, the performance of Model-A is fairly good with computed lattice constants that are slightly smaller than the corresponding experimental values and the computed $\Delta H_{\mathrm{subl}}$ of −16 kcal/mol, which is close to the experimental value of −15.3 kcal/mol taken from ref 30. The space group of the experimental crystal ($P4_2/n$) was maintained during the whole simulation. The computed liquid densities at 300 and 400 K are also in satisfactory agreement with the experimental densities[31] at these temperatures.

Also, for comparison purposes, we executed similar simulations using different force fields. The DREIDING force field,[34] a general purpose force field, has an intermolecular LJ parameter for carbon with three implicit hydrogen atoms. If the methyl parameter of Model-A is replaced by this parameter, the lattice constants become closer to the experimental values, but the interaction energy is weakened, thus resulting in a liquid that possesses too low a density at 300 K (655 kg/m$^3$). On the other hand, at 400 K, these parameters give rise to a vapor phase. Interestingly, the properties computed by using the DREIDING all-atom (Si, O, Me) parameters worsen with respect to the methyl-only simulation, as seen in Table 1, where even at 400 K an amorphous solid structure resulted from the simulation. Further, we also examined the model of Smith et al.,[35] which has been designed for the simulation of poly dimethylsiloxane and uses the rigid body approximation (the molecular geometry of OMCTS used for this model was obtained from a geometry optimization using the Gaussian 03 program). This model, which includes both vdW and electrostatic interaction sites on all atoms, gave good results for the lattice

***Table 1.*** Computed and Experimental Properties for Solid and Liquid OMCTS

| property | Model-A | DREIDING (methyl only) | DREIDING (all atom) | Smith et al. | Model-B | expt. |
|---|---|---|---|---|---|---|
| a (Å) | 15.33 | 16.2 | 15.6 | 16.24 | 15.73 | 16.10[a] |
| b (Å) | 15.33 | 16.2 | 15.6 | 16.24 | 15.73 | 16.10[a] |
| c (Å) | 6.01 | 6.86 | 6.07 | 6.63 | 6.23 | 6.47[a] |
| $\Delta H_{\text{subl}}$ (kcal/mol) | −16.0 | −10.6 | −33.6 | −22.1 | −16.9 | −15.3[b] |
| $\rho_{300}$ (kg/m$^3$) | 1023 | 655 | solid | solid | 948 | 948[c] |
| $\rho_{400}$ (kg/m$^3$) | 801 | vapor | solid | solid | 790 | 830[c] |

[a] Steinfink et al. (ref 19). [b] Osthoff et.al. (ref 32). [c] Palczewska-Tulińska and Oracz (ref 33).

constants, but like for the DREIDING (all atom) case, an amorphous solid was obtained for the liquid state. Hence, it appears that the attractive force between molecules becomes too large when the Si and O terms are added to the intermolecular potential.

In summary, our simulations indicate that Model-A describes quite well the properties of both crystal and liquid phases, thus confirming our assumption that the dominant intermolecular interaction in bulk OMCTS is that among the methyl groups. We also checked how the introduction of atomic charges affects our simulation results. We derived two sets of atomic charges from quantum mechanical calculations performed on the OMCTS molecule, one by fitting the electrostatic potential (ESP) and another by performing a standard Mulliken population analysis of the molecular wave function computed at the MP2/AUG-cc-pVDZ//PBE1PBE/6-31G(d) level of theory. Simulations indicate that including these atomic charges gives slightly shorter lattice constants than those obtained with Model-A, but no specific advantage was seen in spite of the larger computational cost due to the calculation of electrostatic interaction.

## IV. Empirical Refinement

Our simulations indicate that Model-A possesses the essential features for describing the intermolecular interactions in OMCTS. However, because OMCTS is more complex a molecule than methane, there should be room for further refinement of this model. Therefore, we empirically readjusted the two LJ parameters of Model-A so as to reproduce the experimental data. We performed the same simulation as those in the previous section for a number of different parameter sets, among which the best set was chosen so as to minimize the penalty function $F$ expressed as the sum of relative errors on the properties that are listed in Table 1:

$$F = f(a) + f(b) + f(c) + f(\Delta H_{\text{subl}}) + f(\rho_{300}) + f(\rho_{400}) \quad (2)$$

where $f(x)$ is defined as the relative percentage error of $x$:

$$f(x) = \left| \frac{x_{\text{expt}} - x_{\text{calc}}}{x_{\text{expt}}} \right| \times 100 \quad (3)$$

where $x_{\text{expt}}$ and $x_{\text{calc}}$ are the experimental and calculated values of $x$, respectively. Additionally, we excluded the parameter sets that did not produce a vapor phase at 500 K. Following this procedure, the best parameter set was determined as being $\sigma = 3.54$ Å and $\varepsilon = 0.39$ kcal/mol, which we define as Model-B. This pair of parameters is located at the bottom of the contour plot of the penalty function $F$ shown in Figure 3. The coordinates of all the points used to make the contour plot are given in the Supporting Information (Table S1). For the purpose of comparison, the results obtained using Model-B are listed in Table 1.

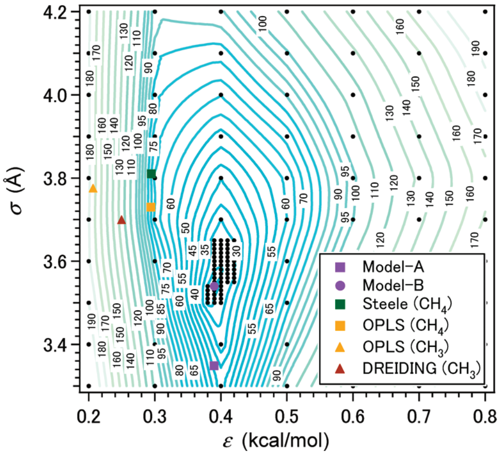


***Figure 3.*** Contour plot of the penalty function $F$. The location of existing LJ parameters of methane ($CH_4$) and the methyl group ($CH_3$) are also plotted in this figure: $\sigma = 3.81$ Å and $\varepsilon = 0.294$ kcal/mol from Steele;[36] $\sigma = 3.73$ Å and $\varepsilon = 0.294$ kcal/mol from OPLS ($CH_4$);[37] $\sigma = 3.775$ Å and $\varepsilon = 0.207$ kcal/mol from OPLS ($CH_3$);[37] $\sigma = 3.7$ Å and $\varepsilon = 0.25$ kcal/mol from DREIDING ($CH_3$).[34]

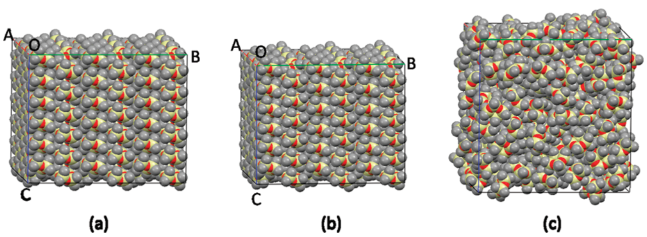


***Figure 4.*** Comparison between (a) experimental and (b) computed OMCTS molecular crystal. (c) Snapshot of the OMCTS liquid. A cubic box was employed for the liquid simulation.

Figure 4a,b shows the experimental and computed crystal unit cells of OMCTS. It is worth noticing that the simulated molecular crystal is characterized by the same type of packing as that of the experimental crystal where the molecules are stacked along the $c$ axis and interact with each other using four methyl groups above and four below the molecular plane. A snapshot of the OMCTS liquid is shown in Figure 4c. Figure 5 shows the calculated radial distribution functions (RDFs) for intermolecular Me−Me pairs in both crystal (at 223 K) and liquid (at 300 and 400 K) simulations.

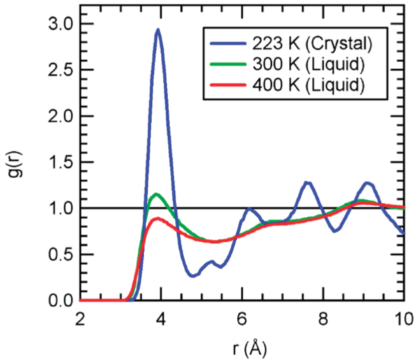


***Figure 5.*** Radial distribution function $g(r)$ of intermolecular methyl−methyl distance derived from the crystal and liquid simulations at different temperatures.

In the RDF of the computed crystal lattice, the first peak appears at 3.9 Å, which compares well with the nearest intermolecular Me−Me distances in the experimental crystal,[19] which are distributed from 3.79 Å to 4.55 Å. Furthermore, there exist four additional peaks at 5.2 Å, 6.2 Å, 7.6 Å, and 9.1 Å, respectively, which reflect the highly ordered packing in the molecular crystal. In comparison, the first peaks of the RDFs of the liquid computed at 300 and 400 K are located near 3.9 Å, but the corresponding heights are considerably smaller than that of the computed crystal. Also, the RDFs of the liquid at these two temperatures show a loss of ordered structure above 5.0 Å.

It is worth mentioning that LJ parameters for methane and methyl groups have been developed by other authors as well. For instance, Jorgensen and co-workers[37] have developed LJ parameters for methane and $CH_3$ parameters for different hydrocarbons which are now part of the OPLS force field, while Steele[36] has proposed LJ parameters for methane which are extensively used in the literature. In principle, such parameter sets could be utilized in the simulation of OMCTS with eq 1 following our strategy (i.e., methyl site only). However, they are located in a zone of the penalty landscape corresponding to larger values of $F$ (see Figure 3).

Finally, we also checked the anisotropy of the intermolecular potential (model-B) as applied to the OMCTS dimer. The potential curves corresponding to the interaction of OMCTS molecules with three different relative orientations, A−C, are plotted in Figure 6. Dimer C with the OMCTS molecules oriented face-to-face possesses the lowest energy ($R$ = 6.0 Å, $E_{\min}$ = −3.69 kcal/mol), the methyl−methyl interactions being maximized. The second minimum is that of dimer A where the OMCTS molecular planes are coplanar to each other ($R$ = 9.0 Å, $E_{\min}$ = −2.70 kcal/mol). The weakest binding is obtained for dimer B where the molecular planes of OMCTS are perpendicular to each other ($R$ = 8.6 Å, $E_{\min}$ = −1.50 kcal/mol). This result is in line with the nonspherical (disk-like) shape of the OMCTS molecule. Hence, the development of such intermolecular potential seems meaningful for the atomistic study of this complex liquid.

## V. Liquid Phase Simulation

In this section, we investigate the performance of Model-B. We performed a series of MD simulations to calculate the

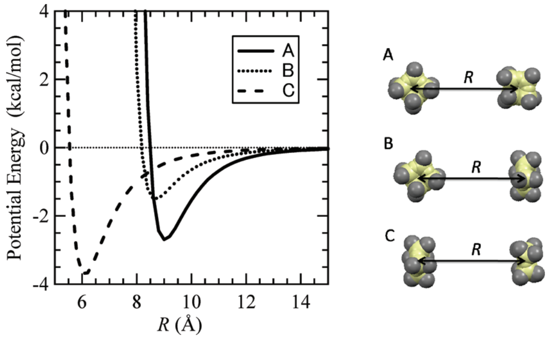


***Figure 6.*** Potential curves of Model-B for three different configurations of the OMCTS dimer. The mean molecular planes of the two molecules are oriented as follows: A, coplanar; B, perpendicular; and C, parallel. The atoms of the siloxane ring define the mean molecular plane of OMCTS. $R$ is the distance between the centers of mass. The relative orientation of OMCTS molecules is kept fixed during the scan.

temperature dependence of the density, diffusion coefficient, and shear viscosity of the liquid phase. The same type of simulation as the liquid case described in section III (with 216 molecules) was executed except that a longer total time step of 15 ns was used. In addition, for $T$ = 300 and 400 K, we also tested a simulation box composed of 640 molecules so as to confirm that the calculated properties do not change when a larger system is employed. The diffusion coefficient $D$ was derived using the Einstein relation (p 60 in ref 12):

$$D = \lim_{t\to\infty} \frac{1}{6Nt} \langle \sum_{i=1}^{N} [\mathbf{r}_i(t) - \mathbf{r}_i(0)]^2 \rangle \tag{4}$$

where $\mathbf{r}_i(t)$ is the center of mass position vector of molecule $i$ at time $t$ and $N$ is the number of molecules in the simulation cell. The quantity in brackets ⟨...⟩ indicates an ensemble average, and here it means taking the average value over different time origins for each time interval $t$. The shear viscosity $\eta$ was calculated by the following equation:[38]

$$\eta = \lim_{t\to\infty} \frac{V}{20k_B Tt} \sum_{\alpha=x,y,z} \sum_{\beta=x,yz} \langle [L^{\alpha\beta}(t) - L^{\alpha\beta}(0)]^2 \rangle \tag{5}$$

where $k_B$ is the Boltzmann constant and $V$ is the volume of the system. $L^{\alpha\beta}$ is defined as

$$L^{\alpha\beta}(t) = \int_0^t P^{\alpha\beta}(\tau)\, d\tau \tag{6}$$

and $P$ is the traceless symmetric part of the stress tensor $\sigma$:

$$P^{\alpha\beta} = \frac{\sigma^{\alpha\beta} + \sigma^{\alpha\beta}}{2} - \frac{\delta^{\alpha\beta}}{3} \sum_{\gamma=x,y,z} \sigma^{\gamma\gamma} \tag{7}$$

where $\delta^{\alpha\beta}$ =1 for $\alpha = \beta$ and 0 for $\alpha \neq \beta$. The stress tensor was calculated as

$$\sigma^{\alpha\beta} = \frac{1}{V}\left[ \sum_{i=1}^{N} \sum_{a\in i} m_a v_a^\alpha v_a^\beta + \sum_i \sum_{j>i} \sum_{a\in i} \sum_{b\in j} (r_a^\alpha - r_b^\alpha) \frac{\partial U}{\partial (r_a^\beta - r_b^\beta)} \right] \tag{8}$$

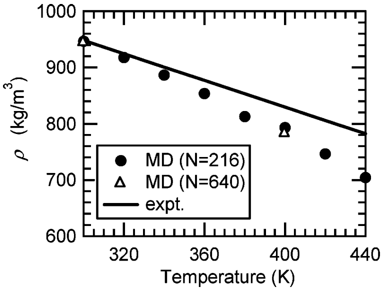


***Figure 7.*** Temperature dependence of liquid density, $\rho$. Experimental curve, $\rho = 1303.79 - 1.18562T$, taken from Palczewska-Tulińska and Oracz.[33] $N$ is the number of molecules in the simulation box.

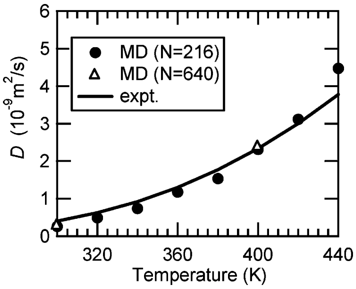


***Figure 8.*** Temperature dependence of diffusion coefficient, $D$. Experimental curve, $D = 10^9 \exp(-14.599 - 2110/T)$, taken from Fischer and Weiss.[40] $N$ is the number of molecules in the simulation box.

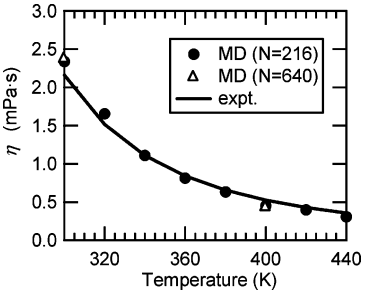


***Figure 9.*** Temperature dependence of shear viscosity, $\eta$. Experimental curve, $\eta = -4.8920 + 1698.7/T$, taken from Palczewska-Tulińska and Oracz.[33] $N$ is the number of molecules in the simulation box.

where $U$ is the potential energy of the system and $m_i$, $r_a^{\alpha}$, and $v_a^{\alpha}$ are the mass, coordinate, and velocity of atom $a$, respectively. The subscript $a \in i$, means the summation on $a$ is taken over molecule $i$.

Equation 5 is the Einstein form of the Green-Kubo relation and is based on the relation derived by Daivis and Evans[39] for an isotropic system:

$$\frac{V}{k_B T} \sum_{\alpha=x,y,z} \sum_{\beta=x,y,z} \int_0^{\infty} \langle P^{\alpha\beta}(t) P^{\alpha\beta}(0) \rangle \, \mathrm{d}t = 10\eta \qquad (9)$$

This relation makes it possible to use all the components of the stress tensor, including the diagonal ones, in the calculation of $\eta$ to enhance the statistical reliability.

The liquid properties computed in the range 300−440 K are compared against the experimental results in Figures 7−9. Figure 7 compares the calculated liquid density ($\rho$) against the experimental values obtained from ref 33. The agreement between calculated and experimental liquid densities is quite good in the low-temperature region around 300 K but degrades by increasing the temperature at above 400 K. Actually, most of the parameter sets tested in the empirical refinement stage underestimated experimental liquid density at 400 K, thus suggesting that additional parameters would be needed to improve this situation. However, the maximum relative error is about 10% (at 440 K), which is good enough if one thinks of the simplicity of the present model, while adding more parameters would make the model computationally less efficient. One possible explanation for the observed deviation of $\rho$ could be related to the rigid body approximation employed here for OMCTS whereas a flexible molecule would better optimize packing.

Figure 8 compares the calculated diffusion coefficient ($D$) against the experimental data from ref 40. The overall agreement with the experiment is good in the whole temperature range of 300−440 K, though the calculated diffusion coefficient increases slightly in the high temperature region. This corresponds to a lower density at these temperatures. Figure 9 compares the calculated shear viscosity ($\eta$) against the experimental data taken from ref 33. The agreement with experiments is good, though the calculated $\eta$ is slightly higher in the low temperature region and lower at the high temperature region. It is noted that the diffusion coefficient and shear viscosity were not included in the parameter optimization process, which again supports that the model correctly captures the essential physics of the intermolecular interaction within liquid OMCTS.

## VI. Conclusions

We developed a new atomistic potential model for the molecular simulation of OMCTS. We could simplify the potential on the assumption that the dominant part of intermolecular interaction in OMCTS is the interaction among methyl groups, as observed in the molecular crystal. Following this approach, the present model possesses only two parameters thereby making it possible to perform an efficient empirical refinement of the potential using the experimental data of both the crystal and liquid phases. Our new model successfully reproduces both the crystal lattice constants and liquid transport properties of OMCTS in a wide range of temperatures, thereby making the large scale atomistic simulation of this molecular liquid possible.

**Acknowledgment.** This research is supported by the Core Research for Evolutional Science and Technology (CREST) program of the Japan Science and Technology Agency (JST). We thank Dr. Bill Smith (Daresbury Laboratory) and the CCP-5 project for a copy of the DL_POLY molecular simulation package. The comments of anonymous reviewers are gratefully acknowledged.

**Supporting Information Available:** A table containing the calculated properties and penalty function $F$ for all the $\varepsilon-\sigma$ pairs explored in this study; a table containing the translational and the first and second order rotational correlation times, $\tau_v$, $\tau_{R1}$, and $\tau_{R2}$ for the liquid phase simulation with Model-B; and the Cartesian coordinates of our rigid-body OMCTS model. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Horn, R. G.; Israelachvili, J. N. *J. Chem. Phys.* **1981**, *75*, 1400–1411.

(2) Israelachvili, J. N. *Intermolecular and Surface Forces*, 2nd ed.; Academic Press: London, 1992.

(3) Klein, J.; Kumacheva, E. *J. Chem. Phys.* **1998**, *108*, 6996–7009.

(4) Demirel, A. L.; Granick, S. *J. Chem. Phys.* **2001**, *115*, 1498–1512.

(5) Kurihara, K. *Prog. Colloid Polym. Sci.* **2002**, *121*, 49–56.

(6) Mizukami, M.; Kusakabe, K.; Kurihara, K. *Prog. Colloid Polym. Sci.* **2004**, *128*, 105–108.

(7) Sakuma, H.; Otsuki, K.; Kurihara, K. *Phys. Rev. Lett.* **2006**, *96*, 046104.

(8) Somers, S. A.; Ayappa, K. G.; McCormick, A. V.; Davis, H. T. *Adsorption* **1996**, *2*, 33–40.

(9) Su, Z.; Cushman, J. H.; Curry, J. E. *J. Chem. Phys.* **2003**, *118*, 1417–1422.

(10) Ayappa, K. G.; Mishra, R. K. *J. Phys. Chem. B* **2007**, *111*, 14299–14310.

(11) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press: New York, 1987.

(12) Haile, J. M. *Molecular Dynamics Simulation: Elementary Methods*; Wiley: New York, 1992.

(13) Stone, A. J. *Science* **2008**, *321*, 787–789.

(14) McCarty, J.; Lyubimov, I. Y.; Gruenza, M. G. *J. Phys. Chem. B* **2009**, *113*, 11876–11886.

(15) Murtola, T.; Bunker, A.; Vattulainen, I.; Deserno, M.; Karttunen, M. *Phys. Chem. Chem. Phys.* **2009**, *11*, 1869–1892.

(16) (a) Lopez, C. F.; Nielsen, S. O.; Srinivas, G.; DeGrado, W. F.; Klein, M. L. *J. Chem. Theory Comput.* **2006**, *2*, 649–655. (b) DeVane, R.; Shinoda, W.; Moore, P. B.; Klein, M. L. *J. Chem. Theory Comput.* **2009**, *5*, 2115–2124.

(17) Spieser, S. A. H.; Leeflang, B. R.; Kroon-Batenburg, L. M. J.; Kroon, J. *J. Phys. Chem. A* **2000**, *104*, 7333–7338.

(18) Peguin, R. P. S.; Kamath, G.; Potoff, J. J.; da Rocha, S. P. *J. Phys. Chem. B* **2009**, *113*, 178–187.

(19) Steinfink, H.; Post, B.; Fankuchen, I. *Acta Crystallogr.* **1955**, *8*, 420–424.

(20) Møller, C.; Plesset, M. S. *Phys. Rev.* **1955**, *46*, 618–622.

(21) Ditchfield, R.; Hehre, W. J.; Pople, J. A. *J. Chem. Phys.* **1971**, *54*, 724–728.

(22) McLean, A. D.; Chandler, G. S. *J. Chem. Phys.* **1980**, *72*, 5639–5348.

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

(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 D.02*; Gaussian, Inc.: Wallingford, CT, 2004.

(25) Li, A. H.-T.; Chao, S. D. *J. Mol. Struct. (Theochem)* **2009**, *897*, 90–94.

(26) Cramer, C. J. *Essentials of Computational Chemistry: Theories and Models*, 2nd ed.; Wiley: Chichester, 2005.

(27) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553–566.

(28) Smith, W. *Mol. Simul.* **2006**, *32*, 933–933.

(29) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(30) Miller III, T. F.; Eleftheriou, M.; Pattnaik, P.; Ndirango, A.; Newns, D. *J. Chem. Phys.* **2002**, *116*, 8649–8659.

(31) Hunter, M. J.; Hyde, J. F.; Warrick, E. L.; Fletcher, H. J. *J. Am. Chem. Soc.* **1946**, *68*, 667–672.

(32) Osthoff, R. C.; Grubb, W. T.; Burkhard, C. A. *J. Am. Chem. Soc.* **1953**, *75*, 2227–2229.

(33) Palczewska-Tulińska, M.; Oracz, P. *J. Chem. Eng. Data* **2005**, *50*, 1711–1719.

(34) Mayo, S. L.; Olafson, B. D.; Goddard, W. A. *J. Phys. Chem.* **1990**, *94*, 8897–8909.

(35) Smith, J. S.; Borodin, O.; Smith, G. D. *J. Phys. Chem. B* **2004**, *108*, 20340–20350.

(36) Steele, W. A. *The Interaction of Gases with Solid Surfaces*, 1st ed.; Pergamon Press: Oxford, 1974; p 56.

(37) Jorgensen, W. L.; Madura, J. D.; Swenson, C. J. *J. Am. Chem. Soc.* **1984**, *106*, 6638–6646.

(38) Mondello, M.; Grest, G. S. *J. Chem. Phys.* **1997**, *106*, 9327–9336.

(39) Daivis, P. J.; Evans, D. J. *J. Chem. Phys.* **1994**, *100*, 541–547.

(40) Fischer, J.; Weiss, A. *Ber. Bunsen—Ges. Phys. Chem.* **1986**, *90*, 896–905.

CT9006053

JCTC Journal of Chemical Theory and Computation

# Search and Characterization of Transition State Structures in Crystalline Systems Using Valence Coordinates

Albert Rimola,† Claudio Marcelo Zicovich-Wilson,*,‡ Roberto Dovesi,† and Piero Ugliengo*,†

*Dipartimento Chimica IFM, NIS Centre of Excellence and INSTM (Materials Science and Technology) National Consortium, University of Torino, Via P. Giuria 7, 10125 Torino, Italy, and Facultad de Ciencias, Universidad Autónoma del Estado de Morelos, Av. Universidad 1001, Col. Chamilpa, 62209 Cuernavaca, Morelos, Mexico*

Received December 19, 2009

**Abstract:** Several tools that allow molecules, polymers, slabs, and crystals to be optimized in valence coordinates as well as a suitable saddle point optimization technique to search for transition state structures for this kind of system have been implemented in the *ab initio* periodic CRYSTAL code. The adoption of these localized coordinate systems largely facilitates the study of chemical processes in periodic systems with atomic connectivity, as occurs in catalytic reactions on zeolites, clathrates, or oxidic surfaces. As a paradigmatic case, the new features have been illustrated to study the proton jump between oxygen atoms of the Brønsted site in the H-chabazite zeolite. The electronic and Gibbs free energy profiles of the most representative proton jump channels have been computed at the B3LYP level, both for a dry H-chabazite as well as in the presence of one $H_2O$ molecule acting as a proton transfer helper. Because of the accuracy allowed by the optimization technique, all stationary points located have been characterized as minima or saddle points by computing the harmonic frequencies and checking, for the latter, that the corresponding transition eigenvectors were in agreement with the selected reaction path. The remarkable agreement between the results with both theoretical and experimental literature data gives credit to the accuracy and robustness of the present implementations in the CRYSTAL code.

## 1. Introduction

The study and characterization of transition state (TS) structures by computational methods is the key step to understanding chemical reactivity. When simulating homogeneous/heterogeneous catalytic processes, it is of paramount relevance to assess the bounty of hypothetical catalysts and to improve their performance through the characterization and design of the activated complexes at the atomic level. Albeit TS optimizations can in principle be performed by means of similar computational strategies of those adopted for minimizations,[1−3] in practice, the former has never become an entirely routine process as it is for the latter. The remarkable difference between both kinds of optimization relies on the quasi-quadratic behavior of the respective optimum domains on the potential energy surface (PES), i.e., while for minima the quadratic basin is large enough to successfully locate a minimum from a wide variety of starting structures, in the TS structures, the basin around a saddle point is much reduced, so that a very good initial guess is required to ensure the convergence through the desired TS by the adopted algorithm. In some way, properly guessing a good starting point still requires a good deal of chemical ingenuity, although several techniques have been proposed to largely help this task.[2,4−9]

The choice of a suitable coordinate system to describe the structures may significantly improve the search of saddle points by enlarging the TS quadratic domains. In molecular cases, a recurrent strategy is based on the use of internal valence coordinates, usually built through Z-matrixes or

* Corresponding author e-mail: piero.ugliengo@unito.it (P.U.); claudio@uaem.mx (C.M.Z.-W.).

† University of Torino.

‡ Universidad Autónoma del Estado de Morelos.

10.1021/ct900680f © 2010 American Chemical Society
Published on Web 03/18/2010

redundant schemes.[1,3] The localized character of these coordinate systems enables a proper description of most of the chemical reactions, involving bond breaking/making processes. This coordinate system, additionally, facilitates a reasonably fast convergence of the optimization technique and permits one to devise straightforward strategies to guess starting structures for the TS optimization.

In periodic calculations, other strategies are customarily preferred to locate TS structures. The most widely used is the so-called climbing image-nudged elastic band (CI-NEB).[10] In this method, the saddle point optimization is substituted by the energy minimization of a supersystem consisting of a set of "images" that sample a path connecting reactants and products which are linked to each other by a kind of "spring forces". The resulting images define the minimum energy path (MEP) of the reaction, the image corresponding to the highest energy being considered the TS. This technique, thus, somehow neglects the localized character of the chemical reactions. The most appealing feature of the CI-NEB method relies on the fact that a suitable starting point for the TS optimization is, in practice, not mandatory, as only the optimized reactant/product structures and the total number of images along the path are needed. Also, the minimization process is, in general, quite stable. However, as a drawback of the method, the nonquadratic character of the overall function considered in the global optimization often slows down the convergence. This is reflected in the known fact that the steepest descent method adopted for the CI-NEB optimization[11] provides directions toward the minima which are far from being conjugate, hence causing slow convergence.[12] In addition, for most periodic codes, the CI-NEB implementation does not allow cell deformations along the MEP because of the complication in the definition of spring forces for lattice vectors. Although extensions of the method have been recently formulated to include cell deformations in the path,[11,13,14] to our knowledge, they have still not been implemented in standard *ab initio* periodic codes.

Other limitations of the bare CI-NEB arise from the coordinate system adopted to define the structure of the images. While for molecular cases the method works unambiguously with Cartesian coordinate systems defined, for each image, on their corresponding center of mass, for periodic systems this is meaningless, as the coordinate system needs to be centered on arbitrary points for each image. Accordingly, depending on this choice, different sets of images can be generated for the same system. For solid state reactions involving bond breaking/making of structures with complex connectivity (as the case of zeolites), the PES defined on the Cartesian coordinates base exhibits complicated shapes which hamper a fast convergence of the CI-NEB method. This may be the reason why the CI-NEB method is rarely used to search for TS in 3D periodic systems, whereas it has been the method of election for studying reactions occurring on 2D surfaces where the above complications are usually absent.

Because of the enormous relevance of heterogeneous catalysis for the modern society, it is even more important to implement the efficient location of TS in the solid state context, the surfaces of many crystals being the places where the catalytically active sites reach their maximum activity.[15] It is worth mentioning that the focuses are not only on the "classical" flat surfaces of compact solids but are also on the internal curved walls of microporous materials like zeolites and metal organic frameworks.[16]

In the present work, we show that the adoption of a system of valence coordinates is not only useful for locating TS in molecules but also in periodic systems with complex connectivity to simplify the study of reaction paths. As shown in what follows, the coordinate system considered here is constituted by a set of redundant internal valence (RIV) coordinates generated according to previous prescriptions on molecular[1,3] and crystalline[17] systems. The adoption of these coordinate systems allows one to consider very straightforward methods for the location of the TS, such as the so-called distinguished reaction coordinate (DRC) one,[18] which is still one of the most used for molecular systems, at least when dealing with few coordinates that dominate the reaction path. In the simplest version of DRC, one degree of freedom, called the distinguished coordinate, is chosen and kept fixed at a sequence of values that are representative of the reaction path, while all the other coordinates are relaxed for each of these values. The maximum-energy geometry along the path is taken as the initial guess for the saddle-point search. Here, the localization of the saddle point requires the calculation of the Hessian matrix of the starting structure. A developing version of the CRYSTAL code[19] is here employed in which the automatic mono- and bidimensional scan along valence coordinates has been implemented.

As anticipated, the behavior of acidic zeolites is of paramount importance because of their central role as catalysts in a large number of key industrial organic transformations.[20−23] The so-called bridging hydroxyl groups belonging to the "Si−O(H)−Al" Brønsted acidic sites exhibit a rather strong acidic proton, as it is well-known that molecules adsorbed in the zeolite cages can easily become protonated in proximity of these acidic sites.[24] Indeed, because proton transfer reactions are key steps in heterogeneously acid-catalyzed reactions, the bridging hydroxyl groups are considered as true catalytic sites. In addition, the mobility of these protons is also important for ion transport in electrolytes, whose potential applications have recently been revealed with the fine-tuning of zeolite-based microfuel cells.[25] In the present work, to focus on a paradigmatic case but relevant from the catalytic point of view, the proton jumps occurring in the acidic chabazite, both in dry conditions and in the presence of one water molecule, have been dealt with. By applying the DRC technique facilitated by describing the problem in RIV coordinates, all the relevant saddle points have been located and characterized by computing both the electronic and free energy barriers of the proton motions. It is worth mentioning that the present implementation allows full relaxation of the cell parameters together with the atomic positions both in the search of the TS domain and in the saddle point optimization. Along this line, it is worth noting that disregarding cell relaxation may (i) introduce some artifacts in the estimation of the reaction energy barriers for localized chemical processes and (ii) hinder the accurate characterization of bulk phase transitions. Both of these facts do highlight the usefulness of the present approach in the study of reactivity in 3D periodic systems.

## 2. Methods

**2.1. Computational Details.** All periodic calculations were carried out with a development version of the ab initio code CRYSTAL06.[19] This code describes the many-electron wave function as a linear combination of crystalline orbitals, which, in turn, are expanded in terms of Gaussian-type functions (GTF), thus allowing one to treat molecules, 1D polymers, 2D surfaces (slabs), and 3D crystals (bulks) with the same level of accuracy.

Both Hartree−Fock (HF), pure PBE[26] and PW91[27] exchange-correlation density functionals (DF), as well as the hybrid PBE0[28] and B3LYP[29,30] DF methods have been used for calculations of the present work with Gaussian basis sets of polarized double-$\zeta$ quality. Details of the adopted GTF basis set are available on the CRYSTAL Web site.[31] Here, for the sake of brevity, only the exponents of the outer shells (in Bohr$^{-2}$) are explicitly given: H, 31G* ($\alpha_s = 0.161$, $\alpha_p = 1.1$); O, 6-31G* ($\alpha_{sp} = 0.27$, $\alpha_d = 0.8$); Si, 66−21G* ($\alpha_{sp} = 0.13$, $\alpha_d = 0.5$); Al, 88−31G* ($\alpha_{sp} = 0.28$, $\alpha_d = 0.47$). To improve the accuracy, some calculations with a polarized triple-$\zeta$ basis set for the atoms of H (311G*, $\alpha_s = 0.10$ and $\alpha_p = 0.75$), O (6-311G*, $\alpha_{sp} = 0.26$ and $\alpha_d = 0.13$), and Si (88−31G*, $\alpha_{sp} = 0.193$, $\alpha_d = 0.61$) have been also carried out.

The Hamiltonian matrix has been diagonalized in 14 reciprocal lattice points (*k*-points), corresponding to a shrinking factor of 3.[32] Tolerances of $10^{-6}$ and $10^{-14}$ were used for the Coulomb and exchange series, respectively.[32] The DFT exchange-correlation contribution is integrated numerically on a grid of points. Radial and angular points of the atomic grid are generated through Gauss-Legendre and Lebedev quadrature schemes. A pruned grid consisting of 75 radial points and a variable number of angular points, with a maximum of 974 angular points in the most accurate integration region (usually named (75, 974)p), has been used.[33,34] The condition for the SCF convergence was set to $10^{-8}$ and $10^{-11}$ Hartree for minima and saddle points, respectively, on the energy difference between two subsequent cycles.

A full relaxation of both lattice parameters and atomic coordinates of the H-chabazite was performed within the *P*1 symmetry. The geometry optimization for minima was performed by means of a quasi-Newton algorithm in which the quadratic step (BFGS Hessian updating scheme) is combined with a linear one (parabolic fit), as proposed by Schlegel.[35] As concerns the TS optimizations, they are performed adopting the method usually referred to as "Eigenvector following" proposed by Simons and Nichols,[36] and the Hessian update has been performed by combining the BFGS and the Murtagh-Sargent approaches[37] in the manner proposed by Bofill.[38] Convergence was tested on the root-mean square (RMS) and the absolute value of the largest component of the gradients and the estimated displacements. The threshold for the maximum force, the RMS force, the maximum atomic displacement, and the RMS atomic displacement on all atoms were set to 0.00045, 0.00030, 0.00180, and 0.00120 au, respectively. By using the same strategy first adopted in Gaussian 80,[39] optimizations were considered complete when the four above conditions were simultaneously satisfied (see ref 40 for specific details on the CRYSTAL06 implementation).

Phonon frequencies of the considered systems have been calculated as the eigenvalues obtained by diagonalizing the mass-weighted Hessian matrix at $\Gamma$ point (point $k = 0$ in the first Brillouin zone, called the central zone). The mass-weighted Hessian matrix was obtained by numerical differentiation (central-difference formula) of the analytical first energy derivatives, calculated at geometries obtained by displacing, in turn, each of the 3*N* equilibrium nuclear coordinates by a small amount, $u = 0.003$ Å. We refer to a recent work[33] for a complete discussion of the computational conditions and other numerical aspects concerning the calculation of the vibrational frequencies at the $\Gamma$ point. Using the optimized geometries and the associated vibrational frequencies, CRYSTAL06 computes the total free energy by correcting the electronic energy by the standard statistical thermodynamics formulas based on partition functions derived from the harmonic oscillator approximations.[41]

**2.2. RIV Coordinates in the CRYSTAL Code and the DRC Strategy.** In the DRC method, a proper initial guess for the direct TS search is needed, so that calculations to pass from the reactant domain to the TS domain of the PES are required. This is achieved by enforcing a number of geometrical constraints as detailed below. In this sense, the most critical issues in the computational implementation of the DRC scheme described above involve the definition of (i) suitable constraints into the optimizations at each fixed point along the reaction coordinate and (ii) a suitable coordinate system for the subspace of the remaining degrees of freedom to ensure the best efficiency of the optimization process.

For the first point, the target is to find a single geometrical parameter (at the moment no more than two parameters are allowed in the present CRYSTAL implementation) suitable to represent the reaction coordinate controlling the TS search. This condition is generally satisfied by choosing internal valence coordinates. Indeed, valence parameter sets, i.e., interatomic distances, angles, and dihedrals, are particularly suitable to describe chemical reactions or phase transitions that essentially involve bonding scheme changes.

For molecules, the valence internal parameters are usually defined by means of the Z-matrix approach. Unfortunately, the resulting Z-matrix coordinate system may not be a good choice for structures that exhibit closed connectivity loops, as the Z-matrix scheme suffers from arbitrariness in the definition of the set, causing slow convergence in the optimization procedure.[1] This is indeed in contrast with the request of point ii described above. For infinite structures like crystals, slabs, or polymers, exploiting symmetry equivalences is essential to reduce the complexity of the system (in principle, infinite) to a degree in which it becomes computationally tractable. Unfortunately, the symmetry constraints (even in the *P*1 case, restricted to translational equivalence) increase dramatically the number of dependencies between internal coordinates, which are formally similar to the closed loops featured by polycyclic molecules, with similar convergence problems. For this reason, the Z-matrix scheme cannot be adopted for geometry optimizations of periodic structures.

A possible solution is to adopt the RIV set of parameters, which allows one to define both the constraints and the coordinate system for the geometry optimization. This coordinate system keeps all the advantages of the valence parameters and additionally reduces the arbitrariness in their definition, allowing a well-balanced description of the structure itself.[1,3,17] The definition of the RIV sets and their implementation in geometry optimizations of molecules[1,3] and crystals[17] have already been reported. Accordingly, the details of the methods will not be repeated here, and only the key differences between the present and previous implementations will be highlighted in the present work.

The first step is to define the atomic connectivity (required to define the RIV coordinates) following the recipe of ref 3. Additionally, in the present implementation, all symmetry equivalences within the RIV set are automatically set up so that an irreducible RIV set that consists of one representative of each symmetry class is kept in memory together with its multiplicity per unit cell, $\mu_t$. A small displacement given in the reference coordinate system, $\delta\mathbf{x}$, can be transformed to $\delta\mathbf{q}$ in the RIV basis set, as

$$\delta\mathbf{q} = \mathbf{B}\delta\mathbf{x} \tag{1}$$

where $\mathbf{B}$ is the Wilson $\mathbf{B}$ matrix whose elements read $B_{ij} = \mu_i\, \partial q_i/\partial x_j$.[1]

For periodic systems, it is customary to adopt as the reference coordinate set both the atomic Cartesian coordinates and the Cartesian lattice vectors, in terms of which the gradient is computed by analytic derivation of the energy.[17] Within this approach, both the external and the nontotally symmetric degrees of freedom constitute the redundant set adopted for geometry optimizations. In the present implementation, however, the atomic part of the reference coordinate set envisages a free set of internal and symmetry adapted linear combinations of the atomic fractional coordinates. The lattice part is a complete free set of symmetry-adapted unit cell elastic distortions. This reference set contains an irreducible number of coordinates free from both external and asymmetric displacements. In order to save both memory and computational time in matrix products, the algorithm coded in CRYSTAL always keeps the gradient defined in terms of this nonredundant reference set in the random access memory. Accordingly, in the following, the reference set $\{x_i\}$ always refers to this nonredundant coordinate system.

The $\mathbf{B}$ matrix is computed by numerical differentiation adopting the central point approximation. The force $\mathbf{f}_\mathrm{q}$ expressed in RIV coordinates is determined from the force $\mathbf{f}_x$ in the reference system as

$$\mathbf{f}_\mathrm{q} = \mathbf{B}^{-}\mathbf{f}_\mathbf{x} \tag{2}$$

where $f_{q_i} = -\partial E/\partial q_i$ and $\mathbf{B}^{-} = \mathbf{G}^{-}\mathbf{B}^{T}$, the superscript "$^{-}$" indicating generalized inverse and $\mathbf{G} = \mathbf{B}^{T}\mathbf{B}$.

To carry out optimizations with constraints as required to explore the PES around the TS, the projector onto the subspace common to the nonredundant and free geometrical subspaces is calculated through[3]

$$\mathbf{P}' = \mathbf{P} - \mathbf{PC}(\mathbf{CPC})^{-}\mathbf{CP} \tag{3}$$

where $\mathbf{P} = \mathbf{G}^{-}\mathbf{G}$ and $\mathbf{C}$ are the projectors onto the nonredundant and the constraint subspaces, respectively. The latter is given in RIV coordinates as

$$\mathbf{C} \equiv C_{ij} = \begin{cases} 1 \text{ if } i = j \text{ and } i \text{ is constrained} \\ 0 \qquad\qquad \text{otherwise} \end{cases} \tag{4}$$

Both the gradient and the Hessian have to be projected out. For the Hessian, the projected matrix $\tilde{\mathbf{H}} = \mathbf{P}'\mathbf{H}\mathbf{P}'$ is diagonalized, and its generalized inverse is computed as

$$[\tilde{H}^{-}]_{ij} = \sum_k T_{ik}[h_k]^{-1}T_{jk} \tag{5}$$

where $T_{ik}$ is the element of the eigenvectors matrix. By identifying the number of redundant degrees of freedom (difference between the dimensions of the RIV and the free subspaces) as $n$, the sum over $k$ in eq 5 includes all eigenvalues of $\tilde{\mathbf{H}}$ with the exception of the $n$ ones exhibiting the lowest absolute value. This selection is performed so as to prevent displacements occurring outside the free subspace defined by $\mathbf{P}'$ and to correct for small numerical errors that derive from the numerical evaluation of the $\mathbf{B}$ matrix.

Once the displacements in the RIV coordinate set are computed, a back-transformation to the Cartesian set is carried out in the iterative manner proposed in ref 3. The numerical calculation of the Hessian matrix for the TS optimization is performed in the reference coordinate system, $\mathbf{H}_x$, in a similar fashion to that considered for the calculation of vibrational modes in CRYSTAL,[33] with the difference that here only the internal totally symmetric displacements are considered. After the construction, the matrix $\mathbf{H}_x$ is transformed to the RIV system according to $\mathbf{H} \equiv \mathbf{H}_q = \mathbf{B}^{-}\mathbf{H}_x(\mathbf{B}^{-})^{T}$.

**2.3. Modeling of the H-Chabazite Structure.** Most of the computational works[42−48] addressed hitherto to predict proton jump barriers in acidic zeolites were carried out by either cluster or embedded calculations, whereas, to our knowledge, no works at a full ab initio level of periodic calculations are available. The disparity of the computed energy barriers (spanning the 12−35 kcal mol$^{-1}$ range, in absence of water) is mainly due to the different methodologies and approaches adopted to model the acidic zeolites. The simplest is based on the cluster approach, which consists of extracting from the infinite solid a finite cluster surrounding the active site. For small clusters, full ab initio calculations have been carried out, whereas for larger clusters, the ONIOM strategy[49] has been adopted. In the latter, the whole systems are divided into different layers, each one being treated at different computational levels: the active site at the highest level of theory, with the rest at the lowest one. Irrespective of these differences, in both procedures, zeolites were treated by means of standard molecular quantum methods so that the lack of long-range effects may cause various pitfalls.[50] The inclusion of long-range effects needs the treatment of the whole solid systems (namely, zeolites

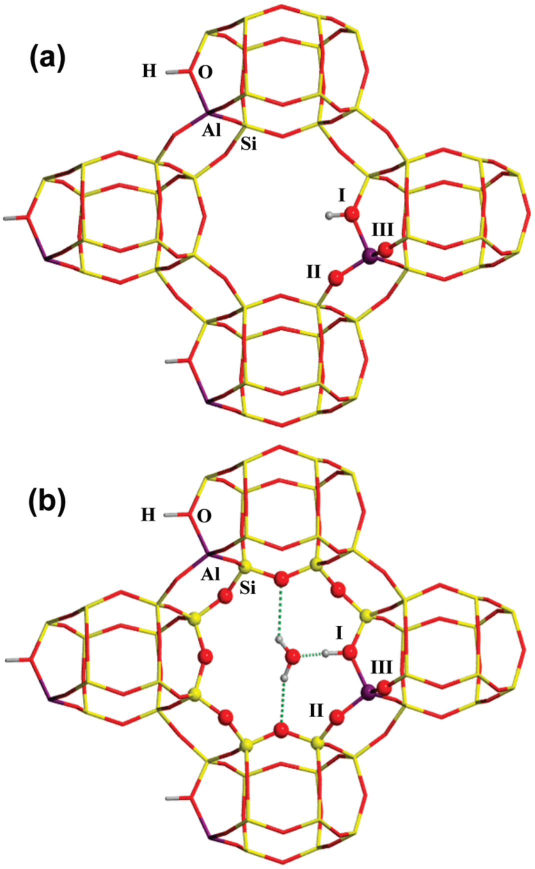


*Figure 1.* H−CHA periodic model used to study the proton jump between sites I, II, and III. Sections: (a) dry H−CHA; (b) H−CHA plus one $H_2O$ molecule.

as infinite solids) via a periodic approach. In the past, periodicity has been exploited either (i) via embedding techniques, such as the QM-Pot method,[51] which partitions the periodic system into two parts, the inner zone, containing the reaction site and treated by quantum mechanics, and the outer zone, described by a properly parametrized classical interatomic potential, and (ii) via full ab initio periodic calculations, ensuring that both the local properties and the long-range effects are treated with the same accuracy. Critical points for the embedding approaches are the absence of charge flux between the inner and outer zones and the difficulty to derive proper interatomic potentials, especially when modeling chemical reactions, whereas the advantage is the freedom to adopt for the inner zone highly accurate quantum mechanical levels (i.e., MP2, CCSD(T), ...), which are not generally available in a fully periodic approach. The latter offers, however, a very clean approach completely free from the previous pitfalls at the expense of being somehow limited to density functional theory as the best level of theory. Some relevant progress has, however, recently been achieved in that respect, as the CRYSCOR program, starting from the Hartree−Fock solution provided by CRYSTAL, can refine the total energy of the crystal at the MP2 level, although at present, it is limited to single point energy evaluation.[52,53]

Considering the above points, in the present work, a periodic model for the H-chabazite has been adopted (shown in Figure 1). It arises from the structure of the pure silica chabazite (CHA),[54] which consists of a network of double six-membered silica rings (hexagonal prisms) connected by four-membered rings. By adopting a Si/Al ratio of 11:1 (one aluminum atom per unit cell), the symmetry of all-silica CHA is reduced from the $R\bar{3}m$ to the $P1$ space group. Charge compensation is achieved by the addition of a proton to one of the four oxygen atoms of the $AlO_4$ tetrahedron (H−CHA). The resulting unit cell (37 atoms) has $HAlSi_{11}O_{24}$ as its chemical formula. The four H−CHA structures exhibit different stabilities, the most favored one corresponding to the proton attached to site I (see Figure 1a).[55] It is well-known that proton jumps from one oxygen to the others may occur, so that the implementation of the DRC method has been tested by computing the energy profile for the proton jumps in H−CHA following the I → II → III → I path (see Figure 1a).[47] For each step, full characterization of minima and the TS has been achieved. Furthermore, the same proton jump route has also been characterized in the presence of one water molecule that acts as a proton transfer helper. Indeed, traces of water in the H−CHA have been demonstrated to significantly reduce the energy barriers.[48,56] For this latter case, the most stable H−CHA/$H_2O$ complex consists of the acidic proton attached to site I engaged in rather strong H-bonds with the $H_2O$ molecule (see Figure 1b).[50]

## 3. Results and Discussion

**3.1. Proton Jump Path I → II with the B3LYP Hamiltonian.** The need to maximize the energy in one (and only one) direction to locate the saddle point on a PES is a very delicate process because too rough TS structures will not converge to the proper final TS. The DRC technique approaches the TS search in two successive steps: (i) defining a structure as close as possible to the TS and (ii) refining this structure to exactly locate the actual TS. As described in the Computational Details section, the present DRC scheme optimizes both the atomic and cell parameters during the search for a TS, at variance with the usual CI-NEB technique in which, despite some recent modifications,[11] the cell part is always kept fixed. In the following, this procedure is described in detail.

*Defining a Geometry Close to the TS.* A geometry close to the TS is defined by a scan calculation along any internal coordinate that may govern the reaction. This scan calculation consists in evolving, step by step and in a controlled way, the selected internal coordinate so as to move from reactants to products by crossing a point of maximum energy. For instance, the proton jump from site I to II implies the breaking and the formation of the O1−H and H−O2 bonds (see Figure 1a), respectively, so that the H···O2 distance may be considered as the internal coordinate that drives the reaction (the so-called reaction coordinate). The scan calculation, based on the H···O2 distance, evolves from the reactant geometry toward the product in many steps along the reaction coordinate; namely, the H···O2 distance shortens from reactants toward products. At each step, the value of the reaction coordinate is frozen while all other internal coordinates are relaxed, so that a “pseudo-optimized” structure is computed. At the end of the scan calculation, a set of

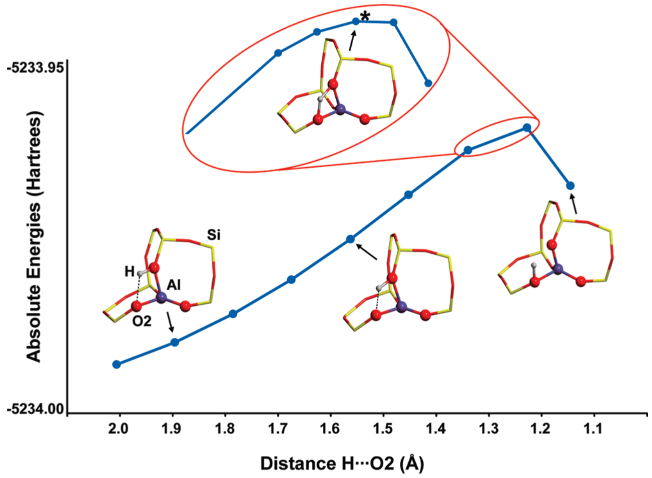


***Figure 2.*** Electronic energy variation during the scan process. The H···O2 distance (restricted reaction coordinate) is frozen at different values, while the remaining internal coordinates are optimized. The red inset zone shows a finer scan with smaller steps close to the energy maximum. The asterisk marks the structure used as the initial guess for the final TS search.

intermediate energy points connecting reactants and products (defined by the various partially optimized structures) is arrived at. Focusing on the present case, the H···O2 distance is scanned from an initial value of 2.0 Å to a final value of 1.1 Å, with 9 partially optimized structures, each differing by a step of −0.1 Å. The energy variation as a function of the DRC is shown in Figure 2. This "distinguished energy profile" exhibits three different zones: (i) a zone where the energy rises moderately as the H···O2 distance shortens (reactant-like zone, the H−O1 bond has not yet been broken), (ii) a zone where the energy is rather high and goes through a maximum (TS-like zone, the proton is almost midway between O1 and O2 positions), and (iii) a zone where the energy decreases (product-like zone, the proton is already bound to O2). The structure at the energy maximum is therefore used to define the initial guess for the next step, i.e., the accurate TS geometry localization. Although not mandatory, a second scan calculation with smaller steps (see inset of Figure 2) around the maximum energy zone (i.e., between 1.3 and 1.2 Å) can be carried out in order to refine the geometry used to secure the initial guess for the TS search.

*Geometry Optimization to a TS.* Starting from the above structure within the TS domain, the Hessian matrix is then computed, either via numerical estimate (very accurate but expensive) or through empirical models (cheaper but less accurate), to ensure that the algorithm will follow the right direction toward the final TS. Irrespective of the way to compute the Hessian matrix, the SCF energy tolerance must however be tightened to $10^{-10}$ to $10^{-11}$ Hartree in order to reach the needed numerical accuracy.

Adopting the above scheme, the TS for the proton transfer from site I to site II has been easily located. In the optimized TS, the H−O2 distance is 1.245 Å, which lies almost between the value for the maximum energy structure (1.26 Å) and the next structure (1.24 Å) in the scan calculation. It is worth mentioning that the same TS structure was arrived at irrespective of the initial guesses which may result from a search based on either a coarse (0.1 Å) or a finer step size (0.02 Å), although for the latter, less optimization cycles are needed to locate the TS. Further details related to the keywords needed to setup both the scan calculation and the TS search are provided in the Supporting Information (SI) as complete CRYSTAL input files.

**3.2. Proton Jump Path I → II with Different Hamiltonians.** To study the dependence of the TS features on the adopted Hamiltonian, the proton jump from site I to II has been computed at HF, PW91, PBE, PBE0, and B3LYP using the same Gaussian basis set reported in the Computational Details section. The structural parameters of the optimized stationary points are reported in Table 1, whereas the reaction energies ($\Delta E_r$) and energy barriers ($\Delta E^{\ddagger}$) are shown in Table 2. Within the same basis set, the computed $\Delta E_r$ values are all rather close to each other as a function of the adopted method, the highest value being for HF (3.3 kcal mol$^{-1}$) compared to density functional values ($\Delta E_r$ = 1.3−2.2 kcal mol$^{-1}$). Similarly, the DF energy barriers, $\Delta E^{\ddagger}$, computed with different DF methods are also very close to each other, lying within a window of 13.6−15.0 kcal mol$^{-1}$, but for that computed at the HF level, which is as high as 30.2 kcal mol$^{-1}$. These values are consistent with the performance of the considered methods in describing proton transfer reactions:[57] the lack of electron correlation in HF yields a dramatic overestimation of the energy barrier, whereas the balanced electron exchange-correlation included in the definition of the DF methods leads to similar $\Delta E^{\ddagger}$ values. The O2−H distances and transition frequency ($\nu^{\ddagger}$) of the TS structures are also dependent on the adopted method, HF systematically providing a too short O2−H distance (1.217 Å) and a too high $\nu^{\ddagger}$ (1801 cm$^{-1}$) compared to the DF methods (O2−H distances lying between 1.233 and 1.245 Å, $\nu^{\ddagger}$ lying between 1085 and 1218 cm$^{-1}$). As quoted in the Introduction, an important issue of the present implementation is its ability to locate the TS by including also the relaxation of the cell parameters. To understand the role that cell parameter relaxation has on the energy barriers, the proton jump has been computed with both PBE and B3LYP functionals by keeping the cell parameters fixed to the values optimized for the reactants. The values reported in Table 2 ("fixed cell" label) show that fixing the cell parameters, while not affecting the thermodynamics ($\Delta E_r$ values), increases the energy barriers $\Delta E^{\ddagger}$ by about 3−4 kcal mol$^{-1}$, reducing the reaction speed by almost 3 orders of magnitude. Accordingly, the $\nu^{\ddagger}$ values for the "fixed cell" cases are all definitely higher than those computed with relaxed cell parameters. Considering that the present TSs involve relatively small molecular aggregates, these results emphasized the relevance of cell relaxation, which will become clearly mandatory for reactions involving bulky reactants.

A comparison of the present results with previous studies allows one to assess whether the actual DRC strategy provides similar potential energy surface features. The closest theoretical work to the present study is the one by Sierka and Sauer,[47] in which the proton jump I → II in the H−CHA was computed by the QM-Pot method, treating the quantum

***Table 1.*** Selected Distances (Å) of the Structures Involved in the Proton Jump from Site I to II, Computed with Different Methods

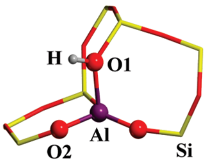
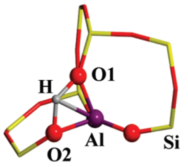
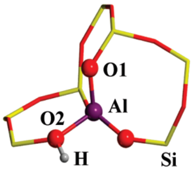
| Structure | Distance | HF | PW91-PW91 | PBE-PBE | PBE0 | B3LYP | B3LYP triple-ζ |
|---|---|---|---|---|---|---|---|
| **I** | O1-H | 0.949 | 0.976 | 0.977 | 0.966 | 0.969 | 0.966 |
|  | O2-H | 2.591 | 2.577 | 2.576 | 2.576 | 2.565 | 2.563 |
| | O1-Al | 1.906 | 1.900 | 1.904 | 1.892 | 1.899 | 1.912 |
| | O2-Al | 1.692 | 1.710 | 1.712 | 1.700 | 1.703 | 1.704 |
| **I → II** | O1-H | 1.227 | 1.251 | 1.250 | 1.210 | 1.222 | 1.233 |
|  | O2-H | 1.217 | 1.233 | 1.237 | 1.238 | 1.245 | 1.239 |
| | O1-Al | 1.791 | 1.808 | 1.812 | 1.826 | 1.830 | 1.837 |
| | O2-Al | 1.802 | 1.816 | 1.819 | 1.816 | 1.820 | 1.825 |
| **II** | O1-H | 3.357 | 3.438 | 3.441 | 3.408 | 3.398 | 3.350 |
|  | O2-H | 0.951 | 0.978 | 0.978 | 0.968 | 0.971 | 0.968 |
| | O1-Al | 1.876 | 1.714 | 1.717 | 1.703 | 1.877 | 1.708 |
| | O2-Al | 1.696 | 1.883 | 1.886 | 1.873 | 1.705 | 1.876 |

***Table 2.*** Electronic Energy Barriers and Reaction Energies, $\Delta E^{\ddagger}$ and $\Delta E_r$ (kcal mol$^{-1}$) for the Proton Jump from Site I to II, Computed with Different Methods[a]

| | $\Delta E$‡ | $\Delta E_r$ | $\nu$‡ |
|---|---|---|---|
| HF | 30.2 | 3.3 | 1801*i* |
| PW91-PW91 | 14.5 | 1.5 | 1089*i* |
| PBE-PBE | 14.4 | 1.3 | 1085*i* |
| PBE-PBE (fixed cell) | 18.4 | 1.6 | 1273*i* |
| PBE0 | 13.6 | 2.1 | 1129*i* |
| B3LYP | 15.0 | 2.2 | 1218*i* |
| B3LYP (fixed cell) | 18.3 | 2.6 | 1334*i* |
| B3LYP/triple-ζ | 17.1 | 3.5 | 1308*i* |
| QM-Pot(B3LYP/T(O)DZP:EVP)[b] | 17.6 | 2.1 | 1151*i* |

[a] Values of proton jump transition frequencies, $\nu^{\ddagger}$ (cm$^{-1}$), are also included. [b] EVP refers to empirical valence bond adopted for the outer zone, see ref 21.

mechanical region at the B3LYP level combining an Ahlrich's Gaussian double-ζ polarized basis set, for H, Si, and Al, and a triple-ζ polarized for O (T(O)DZP). Their energy barrier for the proton jump resulted in 17.6 kcal mol$^{-1}$, 2.6 kcal mol$^{-1}$ higher than the one computed here. To decrease the inconsistency between our standard basis set and that adopted by Sierka and Sauer, the whole DRC was recomputed with the larger basis set described in the Computational Details section (a triple-ζ polarized quality Gaussian basis set) resulting in a $\Delta E^{\ddagger} = 17.1$ kcal mol$^{-1}$, in almost perfect agreement with the value computed by Sierka and Sauer. It is worth noting, however, that whereas the DRC calculations include cell parameter relaxation along the considered points of the potential energy surface, the QM-pot calculations were carried out at fixed cell parameters, causing some inconsistency in the comparison.

**3.3. Proton Jumps in Dry and Wet Conditions.** In this section, the DRC has been adopted to study the proton jump in H−CHA, either in dry conditions or in the presence of one water molecule acting as a proton jump helper. For both cases, the acidic H attached to O1 is the most stable site, and the proton jump has been studied along the following path: I → II → III → I. The computed B3LYP-energy profiles as well as the stationary points are shown in Figures 3 and 4, for the dry and wet cases, respectively.

The intrinsic order of stability (namely, without solvent) of the different $AlO_4$ Brønsted sites resulted in the following order (considering free energies at $T = 298$ K): O1 > O3 > O2. This sequence is identical to those obtained both at the QM-Pot level[47] as well as through full periodic pseudopotential plane-wave calculations.[58] Additionally, experimental measurements revealed that protonation occurs only at sites I and III[59] (note the different numbering of oxygen sites in ref 59), so that our results are consistent with previous data.

In the absence of water, the I → II process has the lowest energy barrier ($\Delta E^{\ddagger}$= 15.0 kcal mol$^{-1}$), whereas the II → III and III → I ones exhibit higher barriers (around 19 kcal mol$^{-1}$). This trend is unchanged also when free energies are considered, albeit the energy barriers are lowered by 4 kcal mol$^{-1}$. The present value of 18.7 kcal mol$^{-1}$ for the $\Delta E^{\ddagger}$ associated to the III → I path is in good agreement with the value of 20.5 kcal mol$^{-1}$ computed by Sierka and Sauer, considering that different basis sets have been used (see the

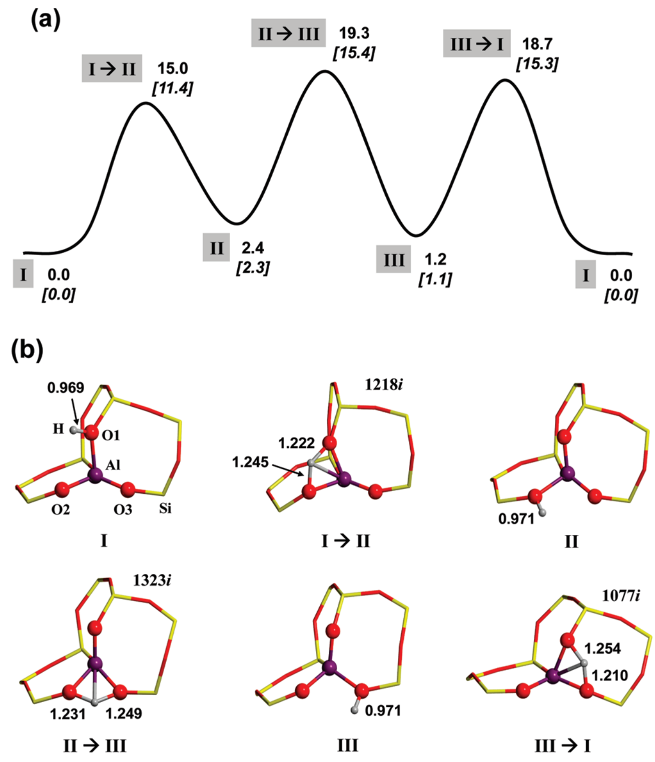


*Figure 3.* (a) B3LYP-energy profile of the proton jump along the I → II → III → I path in dry conditions. Bare values as relative electronic energies, values in brackets as relative free energies at $T = 298$ K with respect to the proton at site I. (b) B3LYP-optimized structures corresponding to stationary points in the energy profile of section a. Bond distances in Å, energies in kcal mol$^{-1}$.

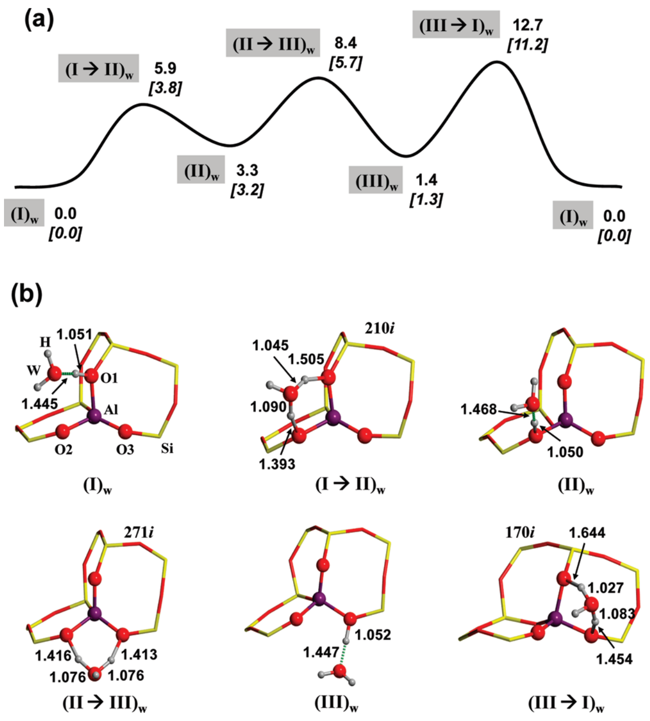


*Figure 4.* (a) B3LYP-energy profile of the proton jump along the I → II → III → I path in the presence of one $H_2O$ molecule acting as a proton transfer helper. Bare values as relative electronic energies, values in brackets to relative free energies at $T = 298$ K with respect to the proton at site I. (b) B3LYP-optimized structures corresponding to stationary points in the energy profile of section a. Bond distances in Å, energies in kcal mol$^{-1}$.

previous discussion for the I → II path). Free cluster calculations reported by Sauer et al.[60] gave $\Delta E^{\ddagger} = 12$ kcal mol$^{-1}$, which is considerably lower than the full periodic B3LYP value. The underestimation of the energy barriers provided by free cluster models compared to more constrained systems was already noticed in a comprehensive study by Fermann et al.,[45] in which results of the proton jumps occurring in H−Y zeolites simulated by free and embedded clusters were compared.

The presence of one water molecule does cause a significant lowering of the energy barriers, indeed acting as a proton helper. Considering only free energies, energy barriers for I → II and II → III paths decrease by 8−10 kcal mol$^{-1}$, whereas for the III → I path, the lowering is only by 4 kcal mol$^{-1}$. These results confirm that, in H−CHA zeolites, the presence of water, even at trace levels, will dramatically alter the proton mobility. The mechanism of proton jump assistance was already observed in zeolites not only for water but also in the presence of other protic solvents, such as methanol and ethanol.[46]

Large differences for the two cases in the $\nu^{\ddagger}$ values have also been predicted: for the water-free H−CHA, the $\nu^{\ddagger}$ value was about 1000 cm$^{-1}$, which reduces to around 200 cm$^{-1}$ (see Figures 3b and 4b, respectively) when water is assisting the proton jump. These values are consistent with the higher geometrical strain (four-membered ring) present in the water-free TS structures compared to the six-membered ring in the presence of one water molecule.

Tuma and Sauer[48] computed the proton jump in H−CHA from site II to III (please note the different numbering scheme of oxygen sites adopted in ref 48) within a QM/QM approach (PBE functional as a lower method, MP2 method as a high-level one). This MP2/DFT scheme gave the proton jump energy barrier of 6.2 kcal mol$^{-1}$ with an Ahlrich's triple-$\zeta$ basis set for O atoms and double-$\zeta$ basis set for the remaining elements. Our own value with the DRC method gives 5.1 kcal mol$^{-1}$ (considering site II as the reference energy asymptote for consistency with ref 48), in good agreement with Tuma and Sauer's value (for basis set effects, see the above discussion). Finally, from the experimental side, $^1$H NMR measurements devoted to the proton exchange rate of different hydrated cation-exchanged (Li, Na, K) CHA allow one to arrive at activation energies ($E_a$) in the range of 10−14 kcal mol$^{-1}$.[61] A direct comparison between our computed barriers and those from the experiment is not straightforward considering the ideality of the adopted model (H−CHA with one $H_2O$ molecule) in contrast to a cation-exchanged CHA with relatively high water loading used in the experiment, and thus the relevance of the comparison should be judged with some extra caution. Our closest result to the experimental results corresponds to the direct proton jump from site I to III ($\Delta E^{\ddagger} = 12.7$ kcal mol$^{-1}$). Nevertheless, a lower barrier (8.4 kcal mol$^{-1}$) is possible via the I → II → III path: notwithstanding, the population of site II will be about 10$^{-3}$ times smaller than that of I, so that experimental barriers may presumably derive from both paths.

## 4. Conclusions

The automatic scan of valence coordinates has been implemented in the CRYSTAL *ab initio* periodic code as a strategy to search and refine saddle point structures, through the DRC strategy. In order to have a consistent coordinate set of valence parameters for periodic systems, the redundant scheme has been adopted.

The performance of the algorithm has been illustrated for the acidic chabazite (H−CHA) zeolite by considering the jump of the acidic proton between different oxygen atoms of the $AlO_4$ tetrahedron. It has been shown that, despite the complexity in the connectivity of the atoms that makes the study of the reactions very difficult when considering Cartesian coordinates, the exploration of the PES around the TS becomes straightforward in terms of valence coordinates, hence allowing the use of quite simple strategies such as the DRC one.

Calculations have been carried out for the proton jump between site I and site II with different Hamiltonians, namely, Hartree−Fock, PBE, PW91, PBE0, and B3LYP hybrid functionals with a double-ζ polarized basis set. As expected, Hartree−Fock overestimates the energy barrier and the transition frequency ($\Delta E^{\ddagger} = 30.2$ kcal mol$^{-1}$ and $\nu^{\ddagger} = 1801$ cm$^{-1}$, respectively) compared to density functional methods ($\Delta E^{\ddagger} = 13.6-15.0$ kcal mol$^{-1}$ and $\nu^{\ddagger} = 1085-1218$ cm$^{-1}$, respectively). It is also shown that cell relaxation during the TS localization significantly influences the accuracy of the results, since barriers to proton jumps computed with fixed cell parameters are by 3−4 kcal mol$^{-1}$ higher than those computed with fully relaxed geometry.

A complete proton jump path, I → II → III → I, has been studied, both in the absence and in the presence of one water molecule which assists the proton jump. Proton jumps in water-free H−CHA exhibit higher energy barriers (among 4−10 kcal mol$^{-1}$) than those computed in the presence of one water molecule, confirming the role of water as a “proton transfer helper”. The free energies of all stationary points have been computed at $T = 298$ K by using the accurate harmonic frequency values provided by the CRYSTAL code, showing values of the free energy barriers somehow lower by 4 kcal mol$^{-1}$ than the purely electronic ones. The accuracy of the present periodic approach to optimize TS structures has been assessed by comparing the electronic energy barriers with those reported in previous theoretical works as well as with the available experimental data, mostly showing very good agreement.

Accordingly, with the present tool, very detailed information of TS structures and free energies can be computed at a reasonable computational cost. The present strategy, moreover, allows one to search for TS structures involving changes in both atomic as well as cell parameters in a very natural way. Relaxing cell parameters may be relevant even for studying localized chemical reactions occurring on adsorbed bulky molecules in zeolite channels, in which expansion or contraction of the silica framework can affect the energy barrier, particularly when dispersive contributions are taken into account.[62] Obviously, this feature is also mandatory for studying phase transition processes, a point which is under investigation in our laboratory. Other issues that deserve further studies concern the implementation of a CI-NEB scheme formulated in terms of RIV coordinates.

**Acknowledgment.** A.R. is indebted to Fundación Ramón Areces for a postdoctoral fellowship. P.U. acknowledges CINECA supercomputing center for a generous allowance of computing time.

**Supporting Information Available:** Input for a scan calculation and input for a TS search calculation. This material is available free of charge via the Internet at http://pubs.acs.org.

### References

(1) Pulay, P.; Fogarasi, G. *J. Chem. Phys.* **1992**, *96*, 2856–2860.

(2) Peng, C.; Schlegel, H. B. *Israel J. Chem.* **1993**, *33*, 449–454.

(3) Peng, C.; Ayala, P. Y.; Schlegel, H. B.; Frisch, M. J. *J. Comput. Chem.* **1996**, *17*, 49–56.

(4) Bell, S.; Crighton, J. S. *J. Chem. Phys.* **1984**, *80*, 2464–2475.

(5) Fischer, S.; Karplus, M. *Chem. Phys. Lett.* **1992**, *194*, 252–261.

(6) Halgren, T. A.; Lipscomb, W. N. *Chem. Phys. Lett.* **1977**, *49*, 225–232.

(7) Hratchian, H. P.; Schlegel, H. B. In *Theory and Applications of Computational Chemistry: The First Forty Years*; Dykstra, C. E., Ed.; Elsevier B. V.: Amsterdam, 2005, pp 195−249.

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

(9) Sholl, D. S.; Steckel, J. A. *Density Functional Theory: A Practical Introduction*; John Wiley & Sons, Inc.: Hoboken, NJ, 2009.

(10) Henkelman, G.; Uberuaga, B. P.; Jónsson, H. *J. Chem. Phys.* **2000**, *113*, 9901–9904.

(11) Caspersen, K. J.; Carter, E. A. *Proc. Natl. Acad. Sci.* **2005**, *102*, 6738–6743.

(12) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes*, 3rd ed.; Cambridge University Press: New York, 2007.

(13) Hennig, R. G.; Trinkle, D. R.; Bouchet, J.; Srinivasan, S. G.; Albers, R. C.; Wilkins, J. W. *Nat. Mater.* **2005**, *4*, 129–133.

(14) Trinkle, D. R.; Hennig, R. G.; Srinivasan, S. G.; Hatch, D. M.; Jones, M. D.; Stokes, H. T.; Albers, R. C.; Wilkins, J. W. *Phys. Rev. Lett.* **2003**, *91*, 025701.

(15) Thomas, J. M.; Thomas, W. J. *Principles and Practice of Heterogeneous Catalysis*; Wiley VCH: Weinheim, Germany, 1996.

(16) Derouane, E. G. *J. Mol. Catal. A: Chem.* **1998**, *134*, 29–45.

(17) Kudin, K. N.; Scuseria, G. E.; Schlegel, H. B. *J. Chem. Phys.* **2001**, *114*, 2919–2923.

(18) Rothman, M. J.; Lohr, J. L.; Ewig, C. S.; Wazer, J. R. V. In *Potential Energy Surfaces and Dynamical Calculations*; Truhlar, D. G., Ed.; Plenum: New York, 1979, pp 653−660.

(19) Dovesi, R.; Saunders, V. R.; Roetti, C.; Orlando, R.; Zicovich-Wilson, C. M.; Pascale, F.; Civalleri, B.; Doll, K.; Harrison, N. M.; Bush, I. J.; D'Arco, P.; Llunell, M. *CRYSTAL06 User's Manual*; University of Torino: Torino, Italy, 2006.

(20) Auerbach, S. M.; Carrado, K. A.; Dutta, P. K. In *Handbook of Zeolite Science and Technology*; Marcel Dekker: New York, 2003.

(21) Corma, A. *Chem. Rev.* **1995**, *95*, 559–614.

(22) Huber, G. W.; Iborra, S.; Corma, A. *Chem. Rev.* **2006**, *106*, 4044–4098.

(23) Meusinger, J.; Corma, A. *J. Catal.* **1996**, *159*, 353–360.

(24) Sauer, J.; Ugliengo, P.; Garrone, E.; Saunders, V. R. *Chem. Rev.* **1994**, *94*, 2095–2160.

(25) Kwan, S. M.; Yeung, K. L. *Chem. Commun.* **2008**, 3631–3633.

(26) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(27) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. *Phys. Rev. B* **1992**, *46*, 6671–6687.

(28) Adamo, C.; Barone, V. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

(29) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

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

(31) http://www.crystal.unito.it/Basis_Sets/Ptable.html (accessed Mar 2010).

(32) Monkhorst, H. J.; Pack, J. D. *Phys. Rev. B* **1976**, *8*, 5188–5192.

(33) Pascale, F.; Zicovich-Wilson, C. M.; Gejo, F. L.; Civalleri, B.; Orlando, R.; Dovesi, R. *J. Comput. Chem.* **2004**, *25*, 888–897.

(34) Pascale, F.; Tosoni, S.; Zicovich-Wilson, C.; Ugliengo, P.; Orlando, R.; Dovesi, R. *Chem. Phys. Lett.* **2004**, *396*, 308–315.

(35) Schlegel, H. B. *J. Comput. Chem.* **1982**, *3*, 214–218.

(36) Simons, J.; Nichols, J. *Int. J. Quantum Chem. Quantum Chem. Symp.* **1990**, *24*, 263–276.

(37) Murtagh, B. A.; Sargent, R. W. H. *Comput. J.* **1970**, *13*, 185–194.

(38) Bofill, J. M. *J. Comput. Chem.* **1994**, *15*, 1–11.

(39) Binkley, J. S.; Whiteside, R. A.; Krishnan, R.; Seeger, R.; Defrees, D. J.; Schlegel, H. B.; Topiol, S.; Kahn, L. R.; Pople, J. A. *Gaussian 80*; Carnegie-Mellon Quantum Chemistry Publishing Unit: Pittsburgh, PA, 1980.

(40) Civalleri, B.; D'Arco, P.; Orlando, R.; Saunders, V. R.; Dovesi, R. *Chem. Phys. Lett.* **2001**, *348*, 131–138.

(41) McQuarrie, D. *Statistical Mechanics*; Harper and Row: New York, 1986.

(42) Fermann, J. T.; Auerbach, S. *J. Chem. Phys.* **2000**, *112*, 6787–6794.

(43) Fermann, J. T.; Blanco, C.; Auerbach, S. *J. Chem. Phys.* **2000**, *112*, 6779–6786.

(44) Sierka, M.; Sauer, J. *J. Chem. Phys.* **2000**, *112*, 6983–6996.

(45) Fermann, J. T.; Moniz, T.; Kiowski, O.; McIntire, T. J.; Auerbach, S. M.; Vreven, T.; Frisch, M. J. *J. Chem. Theory Comput.* **2005**, *1*, 1232–1239.

(46) Ryder, J. A.; Chakraborty, A. K.; Bell, A. T. *J. Phys. Chem. B* **2000**, *104*, 6998–7011.

(47) Sierka, M.; Sauer, J. *J. Phys. Chem. B.* **2001**, *105*, 1603–1613.

(48) Tuma, C.; Sauer, J. *Chem. Phys. Lett.* **2004**, *387*, 388–394.

(49) Dapprich, S.; Komáromi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J. *THEOCHEM* **1999**, *461*−*462*, 1–21.

(50) Solans-Monfort, X.; Sodupe, M.; Branchadell, V.; Sauer, J.; Orlando, R.; Ugliengo, P *J. Phys. Chem. B.* **2005**, *109*, 3539–3545.

(51) Sauer, J.; Sierka, M. *J. Comput. Chem.* **2000**, *21*, 1470–1493.

(52) Maschio, L.; Usvyat, D.; Manby, F. R.; Casassa, S.; Pisani, C.; Schutz, M. *Phys. Rev. B* **2007**, *76*, 075101.

(53) Pisani, C.; Maschio, L.; Casassa, S.; Halo, M.; Schutz, M.; Usvyat, D. *J. Comput. Chem.* **2008**, *29*, 2113–2124.

(54) Díaz-Cabañas, M.-J.; Barrett, P. A. *Chem. Commun.* **1998**, 1881–1882.

(55) Civalleri, B.; Ferrari, A. M.; Llunell, M.; Orlando, R.; Merawa, M.; Ugliengo, P. *Chem. Mater.* **2003**, *15*, 3996–4004.

(56) Chalk, A. J.; Radom, L. *J. Am. Chem. Soc.* **1997**, *119*, 7573–7578.

(57) Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory*; Wiley-VCH: Weinheim, Germany, 2000.

(58) Jeanvoine, Y.; Ángyán, J. G.; Kresse, G.; Hafner, J. *J. Phys. Chem. B* **1998**, *102*, 5573–5580.

(59) Smith, L. J.; Davidson, A.; Cheetham, A. K. *Catal. Lett.* **1997**, *49*, 143–146.

(60) Sauer, J.; Kölmel, C. M.; Hill, J.-R.; Ahlrichs, R. *Chem. Phys. Lett.* **1989**, *164*, 193–198.

(61) Afanassyev, I. S.; Moroz, N. K.; Belitsky, I. A. *J. Phys. Chem. B* **2000**, *104*, 6804–6808.

(62) Tuma, C.; Sauer, J. *Phys. Chem. Chem. Phys.* **2006**, *8*, 3955–3965.

CT900680F

JCTC Journal of Chemical Theory and Computation

# Reactivities of Sites on (5,5) Single-Walled Carbon Nanotubes with and without a Stone-Wales Defect

T. C. Dinadayalane,[†] Jane S. Murray,[‡] Monica C. Concha,[‡] Peter Politzer,[*,‡] and Jerzy Leszczynski[*,†]

*Interdisciplinary Center for Nanotoxicity (ICN), Department of Chemistry and Biochemistry, Jackson State University, 1400 JR Lynch Street, Jackson, Mississippi 39217, and Department of Chemistry, University of New Orleans, New Orleans, Louisiana 70148*

Received December 14, 2009

**Abstract:** The reactivities of various carbon sites on (5,5) single-walled carbon nanotubes (SWCNT) of $C_{70}H_{20}$ with and without a Stone-Wales defect have been predicted computationally. The properties determined include the average local ionization energy $\bar{I}_S(\mathbf{r})$ and pyramidalization angle $\theta_P$ on the surfaces of the bare tubes, the chemisorption energies, bond lengths, stretching frequencies for chemisorbed H and F atoms, and the effects of H and F chemisorption upon the HOMO−LUMO energy gaps. There is a good correlation between the minima of the local ionization energy and the chemisorption energies at different carbon sites, indicating that $\bar{I}_S(\mathbf{r})$ provides an effective means for rapidly and inexpensively assessing the relative reactivities of the carbon sites of SWCNTs. The pyramidalization angle ($\theta_P$), which is a measure of local curvature, also shows a relationship to site reactivity. The most reactive carbon site, identified by having the lowest $\bar{I}_S(\mathbf{r})$ and largest $\theta_P$, is in the Stone-Wales defect region, which also has the least reactive carbon site, having the highest $\bar{I}_S(\mathbf{r})$ and smallest $\theta_P$. The presence of a Stone-Wales defect and also by H and F chemisorption decreased the HOMO−LUMO gap of (5,5) SWCNT.

## Introduction

Linear single-walled carbon nanotubes (SWCNTs), which can be viewed as wrapped-around graphene sheets, are of interest due to their remarkable electrical, mechanical, optical, and chemical properties.[1−4] Insertion of impurities such as ions, metal atoms, and molecules into SWCNTs modifies their band gaps.[5−7] Introducing defects such as Stone-Wales, vacancies, ad-dimers, etc. opens new opportunities for tailoring the electronic properties of SWCNTs.[8−11] Thus, the defect-containing systems and their functionalized forms could be useful for novel applications. Shigekawa and coworkers have demonstrated the creation and destruction of point defects in SWCNTs using scanning tunneling microscopy (STM). This provides a way to precisely control the electronic properties of SWCNTs.[12]

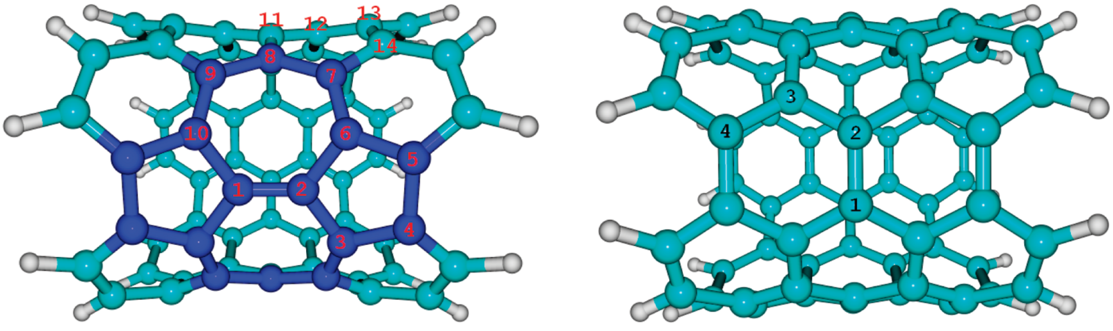
An important nanotube defect is the Stone-Wales defect, which involves four carbon hexagons being replaced by two pentagons and two heptagons coupled in pairs (5−7−7−5; compare **1** and **2** in Figure 1). The Stone-Wales defect is generated by a 90° rotation of a C−C bond in the hexagonal network.[8] The Stone-Wales transformation has an energy barrier of 6−7 eV in a flat graphene sheet and in $C_{60}$.[13−16] Suenaga et al. have shown the first direct imaging of pentagon−heptagon pair defects in a SWCNT by means of high-resolution transmission electron microscopy (HR-TEM) with atomic sensitivity.[17] In-depth theoretical studies of Stone-Wales defects in carbon nanotubes are limited. However, some recent ones have shown the Stone-Wales defect in two different orientations, and its influence on covalent and noncovalent functionalization in selected carbon nanotubes.[18−27] The formation of Stone-Wales defects in the catalytically assisted growth mechanism of SWCNTs has been reported by Charlier et al. using *ab initio* molecular dynamics and tight-binding Monte Carlo simulations.[28]

[*] Corresponding author e-mail: jerzy@icnanotox.org (J.L.); ppolitzer@uno.edu (P.P.).

[†] Jackson State University.

[‡] University of New Orleans.

10.1021/ct900669t © 2010 American Chemical Society
Published on Web 03/10/2010



**1** **2**

***Figure 1.*** Stone-Wales defect (**1**) and defect-free (**2**) armchair (5,5) SWCNTs, $C_{70}H_{20}$. Atom numberings are indicated. The carbon atoms of the Stone-Wales defect region of **1** are shown in blue for clarity.

Pentagon−heptagon pair defects have been observed experimentally and have been reported to play a crucial role in the growth of the nanotube structure.[28] Robinson et al. demonstrated that the chemical sensitivities of SWCNTs can be enhanced significantly by introducing a low density of defects on their sidewalls.[27] Other theoretical studies showed how the defects participate in the chemical sensing behavior of SWCNTs.[25,26,29]

Defect-free SWCNTs, in general, possess few dissimilar carbon atom sites for attachment of functional groups. In contrast, defect-containing tubes have many different sites with varying reactivities, particularly in the region of the defect. Furthermore, the presence of defects changes the local curvatures of SWCNTs, making the carbon atoms in the vicinity of a defect more or less reactive than in the defect-free region or in the pristine tube. Considering the large sizes of nanotubes, it is important to identify and rank the most reactive sites in defect-containing tubes without performing expensive *ab initio* or DFT calculations.

Hydrogen atom chemisorption on the surfaces of SWCNTs has been the subject of both experimental and theoretical studies, since SWCNTs are viewed as a potential means for hydrogen storage.[30−36] Experimental studies by Nikitin et al. on the hydrogenation of SWCNTs with atomic hydrogen showed that it creates C−H bonds, and these C−H bonds can be completely broken by heating to 600 °C.[30] It was reported that hydrogenation of SWCNTs by H-plasma treatment is useful to cut smaller diameter tubes more easily than larger ones.[31] Lu et al. found that chemisorption of H atoms on the exterior surface of the smaller armchair SWCNTs can break the C−C bonds but does not induce unzipping in larger armchair and zigzag SWCNTs.[37] Recently, Stojkovic et al. demonstrated bisection of SWCNT by controlled chemisorption of hydrogen atoms.[32]

Fluorine chemisorption on the surfaces of SWCNTs has evoked experimental interest because fluorine atoms on nanotubes behave as leaving groups and can be readily replaced by nucleophilic agents.[38] Fluorinated nanotubes were characterized by X-ray diffraction and by X-ray photoelectron and Raman spectroscopy.[39] Fluorination followed by pyrolysis of SWCNTs was reported to have "cut" SWCNTs of a range of different lengths.[40] Chemisorption of fluorine atoms on the external surfaces of defect-free SWCNTs has been investigated by various groups,[41−45] but there are no studies involving defect-containing SWCNTs.

One of the objectives of this work has been to examine the reactivities of different carbon atoms in (5,5) armchair SWCNTs with and without the Stone-Wales defect, using the computed average local ionization energy $\bar{I}(\mathbf{r})$, which will be discussed in the next section. The differing reactivities of the carbon atoms will also be analyzed via chemisorptions of H and F atoms on the external surfaces of the tubes.

## The Average Local Ionization Energy

For predicting and interpreting chemical reactivity, which is a local phenomenon that varies from one site to another within a given system, it is essential to have a measure of how readily available, i.e., how strongly held, the electrons are at different sites. It is for this purpose that the average local ionizaton energy, $\bar{I}(\mathbf{r})$, was introduced.[46]

$$\bar{I}(\mathbf{r}) = \frac{\sum_i \rho_i(\mathbf{r})|\varepsilon_i|}{\rho(\mathbf{r})} \qquad (1)$$

In eq 1, $\rho_i(\mathbf{r})$ is the electronic density of orbital $i$, having energy $\varepsilon_i$; $\rho(\mathbf{r})$ is the total electronic density; the summation is over all occupied orbitals.

Within the Hartree−Fock framework, the formalism of the theory plus Koopmans' theorem[47,48] provide support for the approximation $I_i \approx -\varepsilon_i$, where $I_i$ is the ionization energy of the $i$th electron; in density functional theory, Janak's theorem does the same.[49] Thus, $\bar{I}(\mathbf{r})$ can be regarded as the average energy required to remove an electron from the point **r**, the focus being upon the point in space rather than a particular orbital.

While our present interest in $\bar{I}(\mathbf{r})$ is primarily as a guide to reactivity, its significance is more far-reaching. $\bar{I}(\mathbf{r})$ is linked to electronegativity, local kinetic energy density, and local polarizability and hardness. These aspects of it are discussed elsewhere.[50,51]

For interpreting and predicting the reactive behavior of a system, $\bar{I}(\mathbf{r})$ is usually computed on its surface and labeled $\bar{I}_s(\mathbf{r})$. The surface is typically taken to be the 0.001 au (electrons/bohr$^3$) contour of the electronic density $\rho(\mathbf{r})$, as proposed by Bader et al.[52] The local minima of $\bar{I}_S(\mathbf{r})$, designated by $\bar{I}_{S,\min}$, indicate the locations of the least tightly held, most reactive electrons. These are accordingly the preferred sites for electrophilic or radical attack. $\bar{I}(\mathbf{r})$ has indeed proven to be quite effective in analyzing reactive behavior;[46,53−58] for example, it correctly predicts the *ortho/para vs. meta* directing tendencies of benzene substituents, as well as their activation or deactivation of the ring.[46,54] $\bar{I}(\mathbf{r})$ has been utilized for characterizing graphene[55] and nanotube surfaces.[4,59]

## Computational Details

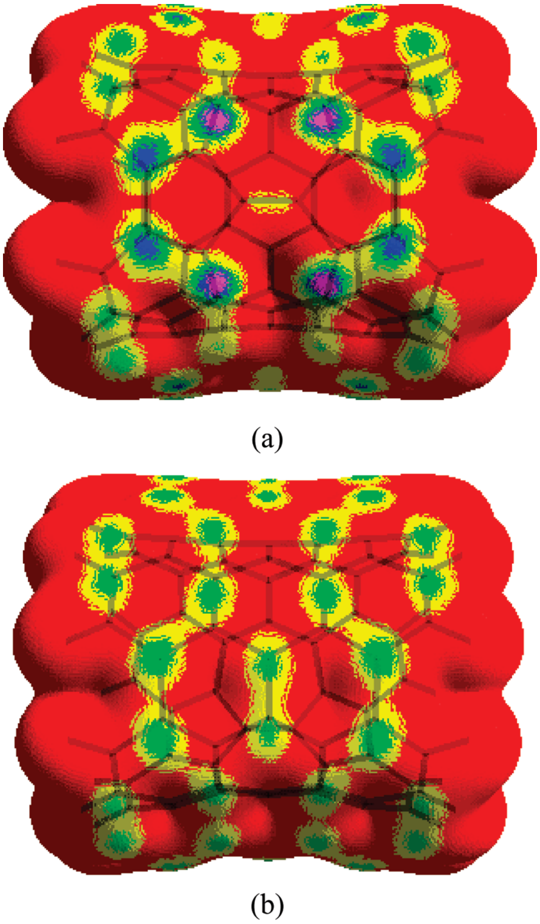
B3LYP/6-31G(d) geometry optimizations were carried out for (5,5) armchair SWCNTs comprised of 70 carbon atoms, both with and without Stone-Wales defects (**1** and **2** in Figure 1). Both ends of the tubes were capped by hydrogen atoms to avoid dangling bonds. The B3LYP/6-31G(d) optimized structures were used to compute the average local ionization energy $\bar{I}_s(\mathbf{r})$ via eq 1 over grids covering both the inner and outer 0.001 au surfaces of the tubes. Due to their large sizes, $\bar{I}(\mathbf{r})$ was calculated at the HF/STO-5G level, which has been found to be quite satisfactory for carbon framework systems.[53,56] For the (5,5) SWCNTs with hydrogen and fluorine atoms bonded at various sites, geometries were optimized and reaction (chemisorption) energies determined with the UB3LYP/6-31G(d) procedure. Vibrational frequency calculations confirmed that all structures correspond to energy minima. All geometry optimizations and frequency calculations were carried out with the Gaussian 03 suite of programs.[60] $\bar{I}_s(\mathbf{r})$ was computed using the HardSurf code.[61]

The numbering of the carbons of tubes **1** and **2** is shown in Figure 1. It should be noted that many sites in each tube are identical by symmetry. For example, C3 in **1** has three identical counterparts. Whatever is said about a particular carbon in the following discussion should be recognized as applying as well to all of its counterparts by symmetry.

## Results and Discussion

Figures 2 and 3 depict the average local ionization energies $\bar{I}_s(\mathbf{r})$ on the outer surfaces of the bare (5,5) SWCNTs **1** and **2**. Table 1 lists the local minima, $\bar{I}_{S,min}$, at various carbon atoms. Both the highest and the lowest $\bar{I}_{S,min}$ are in the defect region of **1**. The lowest $\bar{I}_{S,min}$, predicted to indicate the most reactive carbon, is for C3, which is simultaneously part of five-, six-, and seven-membered rings. The carbon atoms C1 and C2, which form the bond sharing two heptagons, have higher $\bar{I}_{S,min}$ than any other carbons in either the defect-containing (**1**) or the defect-free (**2**) SWCNTs. In **1**, the opposite side of the defect exhibits an $\bar{I}_s(\mathbf{r})$ pattern very similar to that of the defect-free tube **2**; compare Figures 2b and 3.

Table 1 also contains the C−H and C−F bond lengths and the reaction energies ($\Delta E$) for the chemisorption of hydrogen and fluorine atoms at the different carbon sites on the outer sides of **1** and **2**. Each chemisorption is computed to be viable; the F atom chemisorption is more favorable (by 10−12 kcal/mol) than the H atom chemisorption at the corresponding carbon atom site. The C−H distances are all about 1.11 Å, close to the 1.09−1.10 Å that is typical of sp$^3$ carbon,[62] even though the chemisorption energies range from −32.1 to −49.7 kcal/mol. The C−F bond distances are more variable, 1.415 to 1.449 Å, slightly larger than the typical 1.39−1.43 Å.[62] The shortest C−H and C−F bonds and the largest chemisorption energies are at C3 and C14 of **1**. C3 is part of the defect, and C14 adjoins it (Figure 1). In general, Table 1 shows that some of the carbon atoms in the defect


(a)

(b)

***Figure 2.*** Calculated average local ionization energy on the 0.001 au surface of the (5,5) SWCNT of $C_{70}H_{20}$ having a Stone-Wales defect (**1**). Two sides of the tube are shown: (a) the side with the Stone-Wales defect; (b) the side opposite the Stone-Wales defect. Color ranges, in eV: purple, less than 13.3; blue, between 13.3 and 13.5; green, between 13.5 and 14.0; yellow, between 14.0 and 14.7; red, greater than 14.7. Both the highest and the lowest $\bar{I}_s(\mathbf{r})$ are in the defect region.

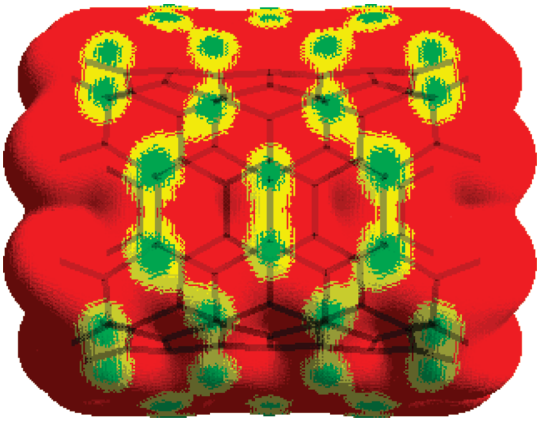


***Figure 3.*** Calculated average local ionization energy on the 0.001 au surface of the defect-free (5,5) SWCNT of $C_{70}H_{20}$ (**2**). Color ranges, in eV: green, between 13.5 and 14.0; yellow, between 14.0 and 14.7; red, greater than 14.7.

***Table 1.*** Computed Properties at Various Carbon Sites of (5,5) Stone-Wales Defect Nanotube **1** and Defect-Free Tube **2**[a]

| | | | | H atom chemisorption | | F atom chemisorption | |
|---|---|---|---|---|---|---|---|
| nanotube | C atom site for H or F chemisorption | $\theta_P$ (deg) | $\bar{I}_{S,min}$ (eV) | C−H distance (Å) | $\Delta E$ (kcal/mol) | C−F distance (Å) | $\Delta E$ (kcal/mol) |
| **1** | C2 | 0.3 | 14.62 | 1.110 | −32.1 | 1.449 | −41.1 |
| | C3 | 7.9 | 13.14 | 1.103 | −49.7 | 1.415 | −59.4 |
| | C4 | 6.4 | 13.32 | 1.107 | −45.6 | 1.428 | −56.4 |
| | C7 | 4.7 | 13.96 | 1.107 | −35.7 | 1.431 | −47.4 |
| | C8 | 2.5 | 14.03 | 1.107 | −33.6 | 1.443 | −45.0 |
| | C11 | 4.6 | 13.93 | 1.106 | −35.7 | 1.436 | −46.2 |
| | C12 | 5.5 | 13.74 | 1.108 | −35.5 | 1.436 | −47.4 |
| | C13 | 6.3 | 13.56 | 1.105 | −42.4 | 1.428 | −54.5 |
| | C14 | 6.5 | 13.45 | 1.103 | −48.5 | 1.419 | −59.5 |
| **2** | C2 | 5.3 | 13.78 | 1.108 | −34.3 | 1.436 | −45.3 |
| | C3 | 5.7 | 13.70 | 1.108 | −35.0 | 1.435 | −46.5 |
| | C4 | 6.0 | 13.55 | 1.105 | −42.8 | 1.427 | −54.6 |

[a] Numbering of carbon atoms is shown in Figure 1.

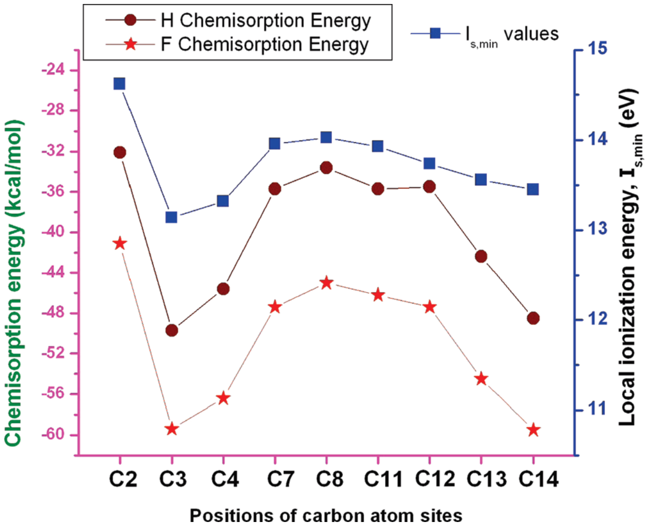


***Figure 4.*** Variation of chemisorption energies (in kcal/mol), calculated at the UB3LYP/6-31G(d) level, for H and F atom chemisorptions at different carbon sites of Stone-Wales defective (5,5) SWCNT (**1**). Plot also shows the minimum values of the average local ionization energy ($\bar{I}_{S,min}$, in eV) at the respective carbon sites.

region are more reactive toward H and F atoms than those in the defect-free tube, while others are less reactive.

Bettinger has investigated computationally, UB3LYP/6-31G(d)//UPBE/3-21G, the reaction energies of fluorine atom additions with (5,5) SWCNTs of various lengths.[45] He found $\Delta E$ to oscillate, being most negative for the fully benzenoid systems. For our tube **2** in Figure 1, he obtained $\Delta E \approx -45$ kcal/mol, very close to our values for C2 and C3 (Table 1).

Figure 4 compares the H and F atom chemisorption energies ($\Delta E$) at various carbon sites on the defect-containing SWCNT (**1**). The trends are very similar, but $\Delta E$ for the fluorine addition is consistently 10−12 kcal/mol larger in magnitude than hydrogen addition. Figure 4 also includes the local minima ($\bar{I}_{S,min}$) of the average local ionization energy $\bar{I}_s(\mathbf{r})$ that are associated with the respective carbon atoms. There is clearly a good correlation between the $\bar{I}_{S,min}$ and the $\Delta E$ for both the H and F chemisorptions. The lower the $\bar{I}_{S,min}$, the larger in magnitude is the chemisorption energy, i.e., the more reactive is the carbon. The only discrepancy is C14, which is more reactive in terms of $\Delta E$ than its $\bar{I}_{S,min}$would predict. $\bar{I}_{S,min}$ and the $\Delta E$ are in agreement concerning the high reactivity of C3, which is shared by five-, six-, and seven-membered rings, and the low reactivity of C2, shared by two seven-membered and one five-membered ring. Both $\bar{I}_{S,min}$ and $\Delta E$ indicate that C3 is the most reactive site and forms the strongest C−H and C−F bonds in the defect portion of the tube. Figure 5 shows the addition of H and F atoms to the most favorable site (C3) of the Stone-Wales defective tube **2**. Atom C2 received considerable theoretical attention since it is involved in the bond rotation to generate the Stone-Wales defect,[21−26] but it is the least reactive among the atoms in the defect region. Figure 4 strikingly demonstrates the general effectiveness of the average local ionization energy in predicting the relative reactivities of different sites of the Stone-Wales defective (5,5) armchair SWCNT. A single calculation deals with the entire surface of a large system, in contrast to the slower and more expensive calculations of $\Delta E$ at different sites.

Table 1 shows that the chemisorptions of H and F atoms at C7, C8, C11, and C12 of the Stone-Wales defect tube **1** have almost the same $\Delta E$ as C2 and C3 of the defect-free system **2**. The reactivity of C13 of **1** is similar to C4 of **2**, both of the carbon atom sites are near the ends of the tubes. The carbon site C14 is a special case, being near an end and also adjoining the defect; its very high reactivity is competitive with that of C3, as already pointed out.

A useful means of characterizing nanotube sites is in terms of their pyramidalization angles, $\theta_P$.[15,19,25,63] This is the angle between the bonds of a given carbon to its three neighbors and the plane defined by those neighbors. The larger is $\theta_P$ at a given site, the greater is the degree of curvature there. Table 1 also lists the $\theta_P$ corresponding to various sites of **1** and **2**, calculated with Haddon's code POAVIT.[64] The carbon atom sites of the Stone-Wales defect region of **1** have a range of $\theta_P$; C2 and C3 possess the lowest and the highest $\theta_P$, respectively. The values of $\theta_P$ for the defect-free tube are intermediate. There are approximate correlations among $\theta_P$, $\Delta E$, and $\bar{I}_{S,min}$. In general, the greater is $\theta_P$, which means the higher the degree of curvature, the greater is the reactivity, for both the Stone-Wales defective and defect-free SWCNTs.

Table 2 provides the energies of the highest-occupied and lowest-unoccupied molecular orbitals (HOMO and LUMO) of **1** and **2**, both bare and with chemisorbed H and F atoms. The HOMO−LUMO energy gaps ($E_{LUMO} - E_{HOMO}$) and the

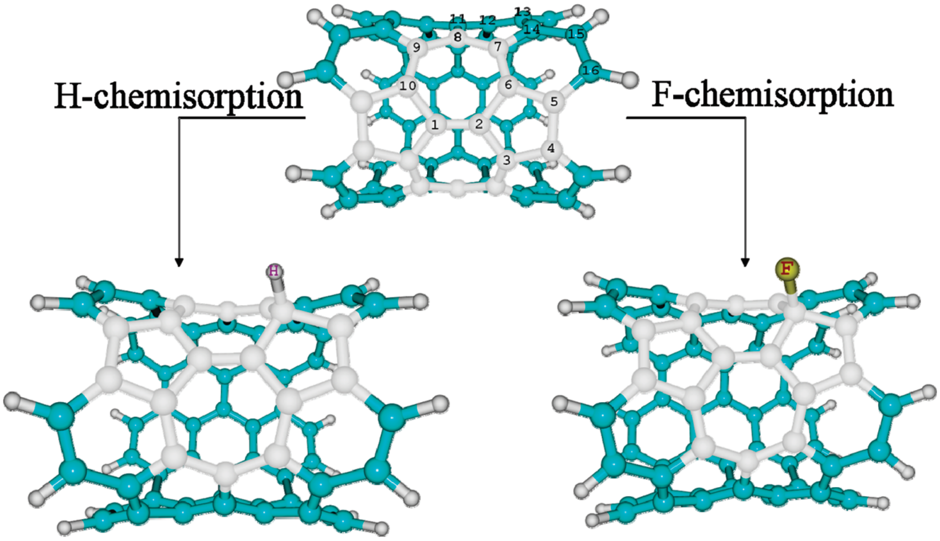


***Figure 5.*** H- and F-chemisorbed Stone-Wales defective SWCNTs obtained by chemisorption of H and F atoms to the most favorable site (C3) in the defect region.

***Table 2.*** Computed Properties of Bare (5,5) Carbon Nanotubes **1** (with Stone-Wales defect) and **2** (defect-free) as well as the Same Tubes with H and F Atom Chemisorbed at Sites Indicated[a]

| | | H atom chemisorption | | | | F atom chemisorption | | | |
|---|---|---|---|---|---|---|---|---|---|
| nanotube | C atom site for H or F chemisorption | HOMO (eV) | LUMO (eV) | H−L gap (eV) | $\nu_{C-H}$ (cm$^{-1}$) | HOMO (eV) | LUMO (eV) | H−L gap (eV) | $\nu_{C-F}$ (cm$^{-1}$) |
| bare **1** | | −4.54 | −2.49 | 2.05 | | −4.54 | −2.49 | 2.05 | |
| **1** | C2 | −4.47 | −2.86 | 1.61 | 2887 | −4.57 | −3.19 | 1.37 | 881 |
| | C3 | −4.54 | −2.64 | 1.90 | 2981 | −4.65 | −2.84 | 1.81 | 1039 |
| | C4 | −4.60 | −2.71 | 1.88 | 2935 | −4.74 | −2.95 | 1.79 | 993 |
| | C7 | −4.64 | −2.90 | 1.74 | 2921 | −4.84 | −3.25 | 1.59 | 973 |
| | C8 | −4.47 | −2.65 | 1.81 | 2910 | −4.54 | −3.08 | 1.46 | 946 |
| | C11 | −4.53 | −2.74 | 1.79 | 2924 | −4.61 | −3.19 | 1.42 | 965 |
| | C12 | −4.63 | −2.63 | 1.99 | 2903 | −4.76 | −3.01 | 1.75 | 966 |
| | C13 | −4.67 | −2.75 | 1.92 | 2944 | −4.80 | −3.08 | 1.73 | 986 |
| | C14 | −4.69 | −2.79 | 1.90 | 2981 | −4.82 | −3.01 | 1.80 | 1029 |
| bare **2** | | −4.51 | −2.30 | 2.21 | | −4.51 | −2.30 | 2.21 | |
| **2** | C2 | −4.47 | −2.77 | 1.70 | 2908 | −4.56 | −3.25 | 1.32 | 972 |
| | C3 | −4.65 | −2.66 | 1.99 | 2907 | −4.83 | −3.09 | 1.74 | 968 |
| | C4 | −4.67 | −2.79 | 1.88 | 2949 | −4.81 | −3.11 | 1.70 | 992 |

[a] Numbering of atoms is shown in Figure 1.

C−H and C−F stretching frequencies $\nu_{C-H}$ and $\nu_{C-F}$ are also included. For a defect-free (5,5) tube with no chemisorbed species, Zhou et al. obtained HOMO and LUMO energies of −4.50 and −2.30 eV,[65] which have been reproduced in our present study for bare tube **2** (Table 2). It has been demonstrated that the HOMO−LUMO gap decreases with increasing tube length, but in an oscillatory manner.[65−67]

Modifications of SWCNTs can produce dramatic effects on electronic properties and can be exploited in the design of novel electronic devices. Therefore, it is important to understand how HOMO−LUMO energy gaps are affected by the Stone-Wales defect and the chemisorptions of H and F atoms at various carbon atom sites. Creating a Stone-Wales defect in a (5,5) SWCNT slightly decreases the HOMO−LUMO energy gap, from 2.21 to 2.05 eV, mainly because of the effect upon the LUMO energy (Table 2). For both defect and defect-free tubes, chemisorptions of H or F atoms generally change both the HOMO and the LUMO energies more negatively, especially the latter. The consequences for the HOMO−LUMO gaps are that they become smaller, particularly when fluorine atoms are chemisorbed. For the Stone-Wales defective SWCNTs, the smallest gap is observed when the H or F is chemisorbed at C2, which was found to be the least reactive site.

The C−H stretching frequencies in Table 2 are all in the 2900−3000 cm$^{-1}$ range, which is consistent with the 2920 cm$^{-1}$ that has been obtained experimentally.[31,68] For comparison, the C−H frequencies in noncyclic alkanes are 2840−3000 cm$^{-1}$, while those in cyclic alkanes and alkenes are 3000−3300 cm$^{-1}$.[69] The highest and the lowest C−H and C−F stretching frequencies are associated with the strongest and the weakest C−H and C−F bonds (as indicated by their $\Delta E$); these are at C3 and C2 sites of the Stone-Wales defect system **1**, respectively.

## Conclusions

We have shown that the average local ionization energy, $\overline{I}_s(\mathbf{r})$, is a good indicator of the relative reactivities of the various carbon atoms of (5,5) armchair SWCNT, both with and without a Stone-Wales defect. In the Stone-Wales defective tube, the most reactive carbon atoms are predicted by $\overline{I}_s(\mathbf{r})$

to be those shared by five-, six-, and seven-membered rings, and the least reactive sites are to be those shared by two seven- and one five-membered ring. The minimum values of $\bar{I}_S(\mathbf{r})$ correlate well with the chemisorption energies of hydrogen and fluorine atom addition at the respective carbon sites. The $\Delta E$ for fluorine addition is about 10−12 kcal/mol more negative than for the hydrogen addition at the corresponding carbon site. Our results indicate that $\bar{I}_S(\mathbf{r})$ is a rapid and inexpensive means for determining the relative reactivities of carbon atom sites of SWCNTs, a single calculation sufficing for the entire surface. The pyramidalization angle $\theta_P$ also exhibits a general relationship to site reactivity. The larger is $\theta_P$, the greater is the local curvature and the more reactive is the carbon atom. The HOMO−LUMO energy gap is decreased by the presence of a Stone-Wales defect and by hydrogen and fluorine chemisorptions.

Various properties were investigated for the chemisorption of H and F atoms, and the results obtained for the defect-free nanotube are generally intermediate in the range obtained for the Stone-Wales defect tube. For example, some carbon atoms in the defect region are more reactive than those in the defect-free system; others are less. The properties of carbon atoms outside of the defect region tend to be similar to those in the defect-free tube. Being near the end of the tube, however, has a modifying influence.

**Acknowledgment.** This work has been supported by the High Performance Computational Design of Novel Materials (HPCDNM) Project funded by the Department of Defense (DoD) through the U.S. Army/Engineer Research and Development Center (Vicksburg, MS), Contract #W912HZ-06-C-0057, and by the Office of Naval Research (ONR), Grant 08PRO2615-00/N00014-08-1-0324. We thank ONR, ERDC, and Mississippi Center for Supercomputing Research (MCSR) for computational facilities.

## References

(1) Saito, R.; Dresselhaus, G.; Dresselhaus, M. S. *Physical Properties of Carbon Nanotubes*; Imperial College Press: London, 1998; pp 1−261.

(2) Ajayan, P. M. *Chem. Rev.* **1999**, *99*, 1787.

(3) Harris, P. J. F. *Carbon Nanotubes and Related Structures*; Cambridge University Press: Cambridge, U. K., 1999; pp 1−283.

(4) Politzer, P.; Murray, J. S.; Lane, P.; Concha, M. C. In *Handbook of Semiconductor Nanostructures and Devices*; Balandin, A. A., King, K. L., Eds.; American Scientific Publishers: Los Angeles, 2006; Vol. 2; pp 215−240.

(5) Zhou, C.; Kong, J.; Yenilmez, E.; Dai, H. *Science* **2000**, *290*, 1552.

(6) Lee, J.; Kim, H.; Kahng, S. J.; Kim, G.; Son, Y. W.; Ihm, J.; Kato, H.; Wang, Z. W.; Okazaki, T.; Shinohara, H.; Kuk, Y. *Nature* **2002**, *415*, 1005.

(7) Yang, S. H.; Shin, W. H.; Kang, J. K. *J. Chem. Phys.* **2006**, *125*, 084705.

(8) Stone, A. J.; Wales, D. J. *Chem. Phys. Lett.* **1986**, *128*, 501.

(9) Smith, B. W.; Luzzi, D. E. *J. Appl. Phys.* **2001**, *90*, 3509.

(10) Orlikowski, D.; Nardelli, M. B.; Bernholc, J.; Roland, C. *Phys. Rev. Lett.* **1999**, *83*, 4132.

(11) Lee, S.; Kim, G.; Ki, H.; Choi, B. Y.; Lee, J.; Jeong, B. W.; Ihm, J.; Kuk, Y.; Kahng, S. J. *Phys. Rev. Lett.* **2005**, *95*, 166402.

(12) Berthe, M.; Yoshida, S.; Ebine, Y.; Kanazawa, K.; Okada, A.; Taninaka, A.; Takeuchi, O.; Fukui, N.; Shinohara, H.; Suzuki, S.; Sumitomo, K.; Kobayashi, Y.; Grandidier, B.; Stievenard, D.; Shigekawa, H. *Nano Lett.* **2007**, *7*, 3623.

(13) Crespi, V. H.; Benedick, L. X.; Cohen, M. L.; Louie, S. G. *Phys. Rev. B* **1996**, *53*, R13303.

(14) Nardelli, M. B.; Yakobson, B. I.; Bernholc, J. *Phys. Rev. B* **1998**, *57*, R4277.

(15) Eggen, B. R.; Heggie, M. I.; Jungnickel, G.; Latham, C. D.; Jones, R.; Briddon, P. R. *Science* **1996**, *272*, 87.

(16) Ewels, C. P.; Heggie, M. I.; Briddon, P. R. *Chem. Phys. Lett.* **2002**, *351*, 178.

(17) Suenaga, K.; Wakabayashi, H.; Koshino, M.; Sato, Y.; Urita, K.; Iijima, S. *Nat. Nanotechnol.* **2007**, *2*, 358.

(18) Dinadayalane, T. C.; Leszczynski, J. Toward nanomaterials: Structural, Energetic and Reactivity Aspects of Single-Walled Carbon Nanotubes. In *Nanomaterials: Design and Simulation*; Balbuena, P. B.; Seminario, J. M. Eds.; Theoretical and Computational Chemistry, Elsevier: Amsterdam, The Netherlands, 2007; Vol. 18, pp 167−199.

(19) Dinadayalane, T. C.; Leszczynski, J. *Chem. Phys. Lett.* **2007**, *434*, 86.

(20) Bettinger, H. F. *J. Phys. Chem. B* **2005**, *109*, 6922.

(21) Yong, S. H.; Shin, W. H.; Lee, J. W.; Kim, S. Y.; Woo, S. I.; Kang, J. K. *J. Phys. Chem. B* **2006**, *110*, 13941.

(22) Wang, C.; Zhou, G.; Liu, H.; Wu, J.; Qiu, Y.; Gu, B.-L.; Duan, W. *J. Phys. Chem. B* **2006**, *110*, 10266.

(23) Akdim, B.; Kar, T.; Duan, X.; Pachter, R. *Chem. Phys. Lett.* **2007**, *445*, 281.

(24) Lu, X.; Chen, Z.; Schleyer, P. v. R. *J. Am. Chem. Soc.* **2005**, *127*, 20.

(25) Andzelm, J.; Govind, N.; Maiti, A. *Chem. Phys. Lett.* **2006**, *421*, 58.

(26) Govind, N.; Andzelm, J.; Maiti, A. *IEEE Sensors J.* **2008**, *8*, 837.

(27) Robinson, J. A.; Snow, E. S.; Badescu, S. C.; Reinecke, T. L.; Perkins, F. K. *Nano Lett.* **2006**, *6*, 1747.

(28) Charlier, J.-C.; Amara, H.; Lambin, Ph. *ACS Nano* **2007**, *1*, 202.

(29) Rivera, J. L.; Rico, J. L.; Starr, F. W. *J. Phys. Chem. C* **2007**, *111*, 18899.

(30) Nikitin, A.; Ogasawara, H.; Mann, D.; Denecke, R.; Zhang, Z.; Dai, H.; Cho, K.; Nilsson, A. *Phys. Rev. Lett.* **2005**, *95*, 225507.

(31) Zhang, G.; Qi, P.; Wang, X.; Lu, Y.; Mann, D.; Li, X.; Dai, H. *J. Am. Chem. Soc.* **2006**, *128*, 6026.

(32) Stojkovic, D.; Lammert, P. E.; Crespi, V. H. *Phys. Rev. Lett.* **2007**, *99*, 026802.

(33) Yang, F. H.; Lachawiec, A. J., Jr.; Yang, R. T. *J. Phys. Chem. B* **2006**, *110*, 6236.

(34) Dinadayalane, T. C.; Kaczmarek, A.; Łukaszewicz, J.; Leszczynski, J. *J. Phys. Chem. C* **2007**, *111*, 7376.

(35) Kaczmarek, A.; Dinadayalane, T. C.; Łukaszewicz, J.; Leszczynski, J. *Int. J. Quantum Chem.* **2007**, *107*, 2211.

(36) Dinadayalane, T. C.; Leszczynski, J. Toward understanding of hydrogen storage in single-walled carbon nanotubes by chemisorption mechanism. In *Practical Aspects of Computational Chemistry: Methods, Concepts and Applications*; Leszczynski, J., Shukla, M. K., Eds.; Springer: New York, 2009; pp 297−313.

(37) Lu, G.; Scudder, H.; Kioussis, N. *Phys. Rev. B* **2003**, *68*, 205416.

(38) Chamsse dine, F.; Claves, D. *Carbon* **2008**, *46*, 957.

(39) Kawasaki, S.; Komatsu, K.; Okino, F.; Touhara, H.; Kataura, H. *Phys. Chem. Chem. Phys.* **2004**, *6*, 1769.

(40) Gu, Z.; Peng, H.; Hauge, R. H.; Smalley, R. E.; Margrave, J. L. *Nano Lett.* **2002**, *2*, 1009.

(41) Chen, Z.; Thiel, W.; Hirsch, A. *ChemPhysChem* **2003**, *4*, 93.

(42) Jaffe, R. L. *J. Phys. Chem. B* **2003**, *107*, 10378.

(43) Kudin, K. N.; Scuseria, G. E.; Yakobson, B. I. *Phys. Rev. B* **2001**, *64*, 235406.

(44) Bettinger, H. F.; Kudin, K. N.; Scuseria, G. E. *J. Am. Chem. Soc.* **2001**, *123*, 12849.

(45) Bettinger, H. F. *Org. Lett.* **2004**, *6*, 731.

(46) Sjoberg, P.; Murray, J. S.; Brinck, T.; Politzer, P. *Can. J. Chem.* **1990**, *68*, 1440.

(47) Koopmans, T. A. *Physica* **1933**, *1*, 104.

(48) Nesbet, R. K. *Adv. Chem. Phys.* **1965**, *9*, 321.

(49) Janak, J. F. *Phys. Rev. B* **1978**, *18*, 7165.

(50) Politzer, P.; Murray, J. S. The average local ionization energy: concepts and applications. In *Theoretical Aspects of Chemical Reactivity*; Toro-Labbé, A., Ed.; Theoretical and Computational Chemistry, Elsevier: Amsterdam, The Netherlands, 2007; Vol. 19, pp 119−137.

(51) Politzer, P.; Murray, J. S.; Bulat, F. A. *J. Mol. Model.* **2010**, accepted.

(52) Bader, R. F. W.; Carroll, M. T.; Cheeseman, J. R.; Chang, C. *J. Am. Chem. Soc.* **1987**, *109*, 7968.

(53) Murray, J. S.; Brinck, T.; Politzer, P. *J. Mol. Struct. (Theochem)* **1992**, *255*, 271.

(54) Politzer, P.; Abu-Awwad, F.; Murray, J. S. *Int. J. Quantum Chem.* **1998**, *69*, 607.

(55) Murray, J. S.; Abu-Awwad, F.; Politzer, P. *J. Mol. Struct. (Theochem)* **2000**, *501−502*, 241.

(56) Peralta-Inga, Z.; Murray, J. S.; Grice, M. E.; Boyd, S.; O'Connor, C. J.; Politzer, P. *J. Mol. Struct. (Theochem)* **2001**, *549*, 147.

(57) Murray, J. S.; Peralta-Inga, Z.; Politzer, P.; Ekanayake, K.; LeBreton, P. *Int. J. Quantum Chem.: Biophys. Quarterly* **2001**, *83*, 245.

(58) Politzer, P.; Murray, J. S.; Concha, M. C. *Int. J. Quantum Chem.* **2002**, *88*, 19.

(59) Politzer, P.; Murray, J. S.; Lane, P.; Concha, M. C. The Remarkable Capacities of (6,0) Carbon and Carbon/Boron/Nitrogen Model Nanotubes for Transmission of Electronic Effects. In *Molecular Materials with Specific Interactions: Modeling and Design*; Sokalski, W. A., Ed.; Challenges and Advances in Computational Chemistry and Physics, Springer: Dordrecht, The Netherlands, 2007; Vol. 4, pp 487−504.

(60) 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.

(61) Sjoberg, P.; Brinck, T. HardSurf program, Ph.D. dissertation, University of New Orleans, New Orleans, LA, 1991, 1993.

(62) Lide, D. R. *Handbook of Chemistry and Physics*, 78th ed.; CRC Press: Boca Raton, FL, 1997.

(63) Haddon, R. C. *Acc. Chem. Res.* **1988**, *21*, 243.

(64) Haddon, R. C. *J. Phys. Chem. A* **2001**, *105*, 4164.

(65) Zhou, Z.; Steigerwald, M.; Hybertsen, M.; Brus, L.; Friesner, R. A. *J. Am. Chem. Soc.* **2004**, *126*, 3597.

(66) Rochefor, A.; Salahub, D. R.; Avouris, P. *J. Phys. Chem. B* **1999**, *103*, 641.

(67) Zurek, E.; Autschbach, J. *J. Am. Chem. Soc.* **2004**, *126*, 13079.

(68) Zhang, G.; Qi, P.; Wang, X.; Lu, Y.; Li, X.; Tu, R.; Bangsaruntip, S.; Mann, D.; Zhang, L.; Dai, H. *Science* **2006**, *314*, 974.

(69) Silverstein, R. M.; Bassler, G. C.; Morrill, T. C. *Spectrometric Identification of Organic Compounds*; Wiley: New York, 1991; pp 103−107.

CT900669T

JCTC Journal of Chemical Theory and Computation

# Decomposing the Energetic Impact of Drug Resistant Mutations in HIV-1 Protease on Binding DRV

Yufeng Cai and Celia A. Schiffer*

*Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, 364 Plantation Street, Worcester, Massachusetts 01605*

Received September 4, 2009

**Abstract:** Darunavir (DRV) is a high affinity ($4.5 \times 10^{-12}$ M, $\Delta G = -15.2$ kcal/mol) HIV-1 protease inhibitor. Two drug-resistant protease variants FLAP+ (L10I, G48V, I54V, V82A) and ACT (V82T, I84V) decrease the binding affinity with DRV by 1.0 and 1.6 kcal/mol, respectively. In this study, the absolute and relative binding free energies of DRV with wild-type protease, FLAP+, and ACT were calculated with MM-PB/GBSA and thermodynamic integration methods, respectively. Free energy decomposition elucidated that the mutations conferred resistance by distorting the active site of HIV-1 protease so that the residues that lost binding free energy were not limited to the sites of mutation. Specifically the bis-tetrahydrofuranylurethane moiety of DRV maintained interactions with the FLAP+ and ACT variants, whereas the 4-amino phenyl group lost more binding free energy with the protease in the FLAP+ and ACT complexes than in the wild-type protease, which could account for the majority of the loss in binding free energy. This suggested that replacement of the 4-amino phenyl group might generate new inhibitors less susceptible to the drug resistant mutations.

## 1. Introduction

The human immunodeficiency virus type 1 (HIV-1, see Abbreviations section at the end for a summary of the abbreviations used in this work) protease is a homodimeric aspartyl enzyme with 99 residues in each chain. The two HIV-1 monomers are bound by nonbonded interactions, with the active site at the interface between the two monomers.[1] The protease processes the viral Gag-Pol polyprotein, yielding the structural proteins and enzymes critical for the maturation of infectious viral particles.[2] Thus, HIV-1 protease has been a major target for structure-based drug design. Nine protease inhibitors have been approved by the Food and Drug Administration (FDA) for HIV therapy, effectively decreasing the mortality rate of HIV/AIDS patients.[3] These FDA-approved HIV-1 protease inhibitors, developed at least in part using structure based drug design, are competitive inhibitors.[2] Unfortunately, exposure to protease inhibitors selects for viruses that have acquired drug resistance mutations in protease due to the high replication rate of HIV-1 and to lack of a proofreading mechanism in its reverse transcriptase. These drug-resistant protease variants lose their high binding affinity to the inhibitors, while maintaining enough enzyme activity for the virus to propagate.[4]

To understand the basis for these changes in drug-resistant proteases, over 200 crystal structures of HIV-1 protease variants have been solved in the past 25 years. Changes in affinity due to drug resistant mutations and thus the thermodynamics of binding can be measured by isothermal titration calorimetry.[5,6] Comparison between the structures of wild-type and drug-resistant variant proteases in complex with inhibitors partially elucidates how specific protease mutations decrease protease−inhibitor binding affinity.[7,8] However, elucidating the critical components of the binding affinity quantitatively from the structural data still remains a challenge. Free-energy simulations,[9–15] in principle, can aid in elucidating these components of the binding affinities to particular atomic interactions.

Among these computational methods, free-energy perturbation (FEP) and thermodynamic integration (TI) methods, which are derived from statistical mechanics,[12,16–21] are mostly used with the thermodynamic cycle to calculate relative binding free energy changes in similar systems. The

* Corresponding author e-mail: celia.schiffer@umassmed.edu.

10.1021/ct9004678
Published on Web 03/04/2010

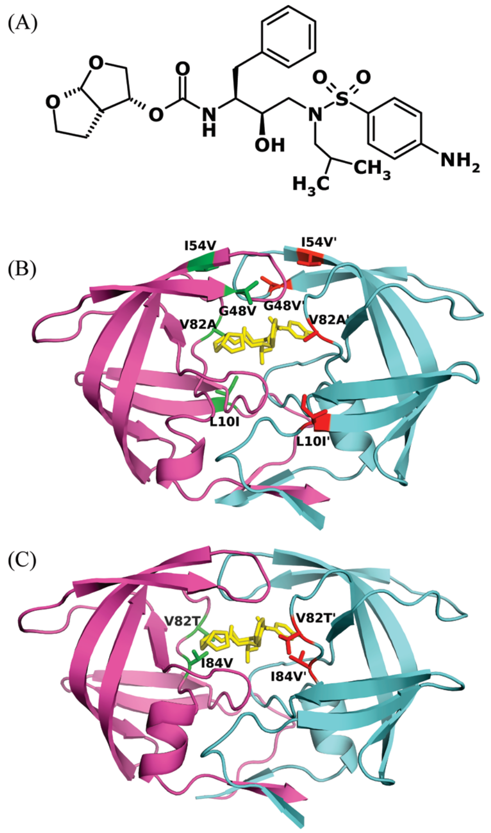


***Figure 1.*** (A) Chemical structure of DRV. (B) Crystal structure of protease variant FLAP+−DRV complex 3EKT.[66] DRV is colored yellow. The side chains of the mutated residues Ile10, Val48, Val54, and Ala82 are displayed and colored red or green. (C) Structure of protease variant ACT−DRV complex 1T7I.[5] DRV is colored yellow. The side chains of the mutated residues Thr82 and Val84 are displayed and colored red or green.

molecular-mechanics Poisson−Boltzmann surface area (MM-PBSA) method combines molecular mechanics and the continuum solvation model.[13,22–26] Solvation properties can be described by the Poisson−Boltzmann (PB) or generalized Born (GB) equation. This method is reliable and applicable to calculating absolute binding free energy change associates with biomolecular recognitions. To achieve a better match with experimental data, the MM-PB/GBSA method is usually supplemented by rough entropy estimation. Free-energy calculation methods provide a way to estimate the binding free energy of inhibitors with different protease variants, allowing computational screening of lead compounds in rational drug design. Furthermore, the calculation results can be further analyzed, e.g., for free energy decomposition, to provide information about affinity changes due to specific kinds of interaction on an atomic level, which could not be determined by experimental methods.[13,22,27,28]

The HIV-1 protease inhibitor, Darunavir (DRV, formerly known as TMC114; Figure 1A) has recently been approved by the FDA.[29] This second-generation protease inhibitor, which was developed after extensive effort in rational drug design,[30] binds the most tightly to the protease of all known inhibitors ($K_d = 4.5 \times 10^{-12}$ M).[5] Nonetheless, DRV still loses affinity to drug resistant variants of HIV-1 protease.[5] In this study, the binding of DRV was investigated with wild-type HIV-1 protease and two drug-resistant variants: FLAP+ (Figure 1B) with L10I, G48V, I54V, and V82A, which are a combination of flap and active site mutations, and ACT (Figure 1C) with V82T and I84V, which are active site mutations. Each of these three systems was analyzed in three parallel 20 ns molecular dynamics (MD) simulations using initial coordinates from their crystal structures. In these MD simulation trajectories, the MM-PBSA and MM-GBSA methods were applied to calculate changes in binding free energy, which were compared with ITC results. The classical TI method was also used to calculate and compare differences in binding free energy of the DRV-ACT and DRV-WT complexes. The accuracy, convergence, and reproducibility of the calculated results have been compared and discussed. The MM-PB/GBSA correctly predicted the order of binding affinity of DRV-WT, DRV-Flap+, and DRV-ACT. The TI calculation result is in good agreement with the experimental data. Free energy component analysis is performed to elucidate the mechanism for resistance of FLAP+ and ACT to DRV. The free energy decomposition study results show that the *bis*-THF group of DRV has maintained its favorable van der Waals (vdW) contact with the protease even in the drug resistant variants. Understanding how the protease mutates to decrease its binding affinity with a very high affinity inhibitor will contribute to developing better strategies to design protease inhibitors.

## 2. Methods

**2.1. MD Simulation with the Program Sander in the AMBER 8 Package.** The initial coordinates of the DRV−WT, DRV−FLAP+, and DRV−ACT protease complexes were taken from each of their respective cocrystal structures 1T3R,[5] 3EKT, and 1T7I.[5]

Molecular dynamics simulations were performed using the program Sander in the MD simulation package AMBER 8.[31] For the standard protease residues, the atomic partial charges, van der Waals parameters, equilibrium bond lengths, bond angles, dihedral angles, and their relative force constants were taken from the AMBER database (ff03).[32] For DRV parameters, the van der Waals parameters, equilibrium bond lengths, bond angles, dihedral angles, and force constants were taken from the General AMBER Force Field database.[33] The partial charges of inhibitor atoms were obtained as follows. First, the coordinates of the DRV atoms were taken from the 1T3R crystal structure and the missing hydrogen atoms added by the program Quanta. Second, the geometry of the resulting structure was optimized with the (HF)/6-31G* basis set by the Gaussian 03 package.[34] Finally, the resulting electrostatic potential was used in the RESP[35] module of the AMBER 8 package to derive the atomic partial charges of the inhibitor.

The explicit solvent model was applied to all systems. Each structure was solvated with the TIP3P water cubic box to allow for at least 8 Å of solvent on each face of the protease. The vdW dimensions for the protease were 44 by 35 by 59 Å. The dimensions of the final periodic box were 63 by 55 by 78 Å. The simulation system had approximately

7000 water molecules, and six $Cl^-$ counterions were added to balance the charge of the system.

A three-step energy minimization process with the steepest descent method was used to allow the system to reach an energetically favorable conformation. In the first energy minimization step, all the heavy atoms of the protease were restrained with a harmonic force constant of 10 kcal mol$^{-1}$ Å$^{-2}$. In the second step, only the backbone nitrogen, oxygen, and carbon atoms were restrained. The strength of the restraint was maintained as 10 kcal mol$^{-1}$ Å$^{-2}$. In the third step, the restraint was turned off, and all atoms were allowed to move. Each of the three steps had 2000 cycles. The temperature of the energy-minimized system was then gradually raised from 50 K to 300 K in the NVT ensemble. Initial velocities were assigned according to the Maxwellian distribution, and random seeds were assigned with three different values to generate nine simulations, three parallel simulations for each of the WT−DRV, FLAP+−DRV, and ACT−DRV systems. In the thermalization process, heavy atoms were restrained with a harmonic force constant of 10 kcal mol$^{-1}$ Å$^{-2}$. The whole process was 50 ps (50 000 steps, each of which was 1 fs). A 50 ps equilibration was then performed in the NPT ensemble without restraining heavy atoms. In the subsequent sampling MD simulations, each step was 2 fs, and the total simulation was 20 ns. For the thermalization, equilibration, and sampling simulations, the SHAKE algorithm[36] was applied to constrain all hydrogen atoms.

**2.2. MM-PB/GBSA Method.** For the protease−ligand system, the binding free energy change is represented by

$$\text{Protease} + \text{Inhibitor} \xrightarrow{\Delta G_{\text{binding}}} \text{Complex}$$

and

$$\Delta G_{\text{binding}} = \Delta G_{\text{MM}} - T\Delta S + \Delta G_{\text{PB/GB}} + \Delta G_{\text{NP}}$$

where

$$\Delta G_{\text{MM}} = \Delta G_{\text{bond}} + \Delta G_{\text{angle}} + \Delta G_{\text{dihe}} + \Delta G_{\text{vdW}} + \Delta G_{\text{ele}}$$

$$\Delta S = \Delta S_{\text{translational}} + \Delta S_{\text{rotational}} + \Delta S_{\text{vibrational}}$$

The molecular mechanical energy $\Delta G_{\text{MM}}$ is the estimated free energy change associated with the binding process in the gas phase. $\Delta G_{\text{MM}}$ was calculated by standard force field functions and parameters. Depending on the type of interaction, $\Delta G_{\text{MM}}$ has two kinds of energetic terms: bonded and nonbonded. The bonded term includes terms representing bond stretching energy ($\Delta G_{\text{bond}}$), angle vibrational energy ($\Delta G_{\text{angle}}$), and dihedral angle torsion energy ($\Delta G_{\text{dihedral}}$). The nonbonded term includes terms representing the van der Waals interaction energy ($\Delta G_{\text{vdW}}$) and electrostatic interaction energy ($\Delta G_{\text{ele}}$).

The polar component of the solvation free energy, represented by $\Delta G_{\text{PB/GB}}$, can be calculated either by solving the Poisson−Boltzmann equation (PB method) or the generalized Born equation (GB method). The nonpolar component of the solvation free energy is represented by $\Delta G_{\text{NP}}$. The sum of $\Delta G_{\text{PB/GB}}$ and $\Delta G_{\text{NP}}$ estimates the free energy change associated with molecules entering solvation from the gas phase. The GB calculation was done using the model developed by Onufriev et al.[37,38] The PB calculation was done with the AMBER 8 numerical PB solver.[39] The solute dielectric constant is 1.0, and the solvent dielectric constant is 80.0. $\Delta G_{\text{NP}}$ was calculated by the LCPO (linear combinations of pairwise overlaps) method, which is linearly dependent on the solvent access surface area: $\Delta G_{\text{NP}} = 0.0072 \times \text{SASA}$.[40] The entropy was calculated by normal-mode analysis using the AMBER 8 NMODE module.[26,41] For every 20 ps of the 20 ns MD simulation trajectory, a snapshot of the protease and inhibitor was taken removing the solvent and counterions. The total number of the atoms for each of the three systems DRV−WT, DRV−FLAP+, and DRV−ACT were 3203, 3209, and 3203, respectively. Altogether, 1000 frames were used for the MM-PB/GBSA calculations. The time-consuming entropy calculations were performed on 100 frames.

**2.3. Thermodynamic Integration Method.** When studying drug-resistant protease mutants, the binding free energy relative to wild-type protease is even more important than the absolute binding free energy. The thermodynamic integration method[42,43] was applied to the protease−inhibitor system to compute the free-energy difference between different states of the system. From statistical mechanics, the Gibbs free energy ($G$) can be calculated from the partition function $Q$ as follows:

$$G = -RT \ln Q \tag{1}$$

The partition function can be expressed as the integral of the system's Hamiltonian function $H(r,p)$. After a coupling parameter, $\lambda$, is introduced into the Hamiltonian, $Q$ can be expressed as

$$Q = \int \int \mathrm{d}r\, \mathrm{d}p \exp(-H(r,p,\lambda)/RT) \tag{2}$$

From eqs 1 and 2, the derivative of $G$ with respect to $\lambda$ is

$$\frac{\mathrm{d}G}{\mathrm{d}\lambda} = \frac{\int \int \frac{\mathrm{d}H(r,p,\lambda)}{\mathrm{d}\lambda} \mathrm{e}^{-H(r,p,\lambda)/RT}\, \mathrm{d}r\, \mathrm{d}p}{\int \int \mathrm{e}^{-H(r,p,\lambda)/RT}\, \mathrm{d}r\, \mathrm{d}p} = \left\langle \frac{\mathrm{d}H(r,p,\lambda)}{\mathrm{d}\lambda} \right\rangle_{\lambda} \tag{3}$$

and

$$\Delta G = \int_0^1 \left\langle \frac{\partial H(\lambda)}{\partial \lambda} \right\rangle_{\lambda} \mathrm{d}\lambda \tag{4}$$

Equation 4 is the master equation of the thermodynamic integration method. When applying this equation to the protein−ligand system, the kinetic component of the Hamiltonian can be neglected. Thus, the $\lambda$-coupling force field function $V(\lambda,r)$ was used to replace the Hamiltonian. The $\lambda$ was chosen such that, when it equals zero, the force field function $V(0)$ and its relative parameters were correlated with the wild-type protease. When $\lambda = 1$, $V(1)$ and its parameters were correlated with the mutant protease. The numerical estimation of eq 4 was

$$\Delta G \approx \sum_{i=1}^{n} w_i \left\langle \frac{\partial V}{\partial \lambda} \right\rangle_{\lambda i} \qquad (5)$$

The $\lambda$ values and their relative weights (Table 5) were assigned from the Gaussian quadratic formula.

Directly calculating thermodynamic integration from the unbound to the bound state is not feasible. The thermodynamic cycle below was used since $G$ represents a state function and is independent of the path.

$$\begin{tikzcd}
\text{Protease}_{\text{WT}} + \text{Inhibitor} \arrow[r, "\Delta G_{\text{WT}}"] \arrow[d, "\Delta G_{\text{apo}}"] & \text{Complex}_{\text{WT}} \arrow[d, "\Delta G_{\text{complex}}"] \\
\text{Protease}_{\text{ACT}} + \text{Inhibitor} \arrow[r, "\Delta G_{\text{ACT}}"] & \text{Complex}_{\text{ACT}}
\end{tikzcd}$$

As shown above, instead of calculating the free energy changes $\Delta G_{\text{WT}}$ and $\Delta G_{\text{ACT}}$ associated with the chemical reaction path, the $\Delta G_{\text{apo}}$ and $\Delta G_{\text{complex}}$ through the "alchemical" path[44−46] were calculated.

Thus, the drug-resistant mutant's loss of binding free energy compared to the wild-type protease was represented by

$$\Delta\Delta G = \Delta G_{\text{ACT}} - \Delta G_{\text{WT}} = \Delta G_{\text{complex}} - \Delta G_{\text{apo}} \qquad (6)$$

The thermodynamic integrations were carried out in the Sander module of the AMBER 8 package.[47,48] The wild-type and mutant proteases have different numbers of side chain atoms on the mutated residues. To keep the same number of atoms in the initial and final states, we perturbed the extra atoms to dummy atoms, which had no nonbonding interactions with the rest of the system. For the ACT mutant, both mutated residues (V82T and I84V) have fewer atoms than the wild type. Thus, the perturbation was done from WT to ACT (Figure 2).

The DRV−WT crystal structure 1T3R was used to generate the coordinates file for the calculation of the $\Delta G_{\text{complex}}$. For the calculation of $\Delta G_{\text{apo}}$, two sets of coordinates were used. One was from the unbound wildtype protease crystal structure 1HHP. The other one was the protease atoms coordinates from the WT−DRV complex crystal structure 1T3R with the inhibitor of DRV deleted from the set of coordinates. The three-step energy minimization was performed as described above. The structure was then thermalized and pre-equilibrated with a harmonic restrained force, and the $\lambda$ value was 0.5. During the thermalization, different random seed values were assigned to create parallel calculations as controls. The pre-equilibrated structure was then sampled at 12 $\lambda$ values, see Table S1 (Supporting Information). The pre-equilibrated structure was then used to start the 12 independent simulations with different corresponding $\lambda$ values (Table S1). The time steps were 1 fs, and the time for the calculation at each $\lambda$ value was 2 ns. Thus, the total sampling time for each alchemical free energy change calculation was 24 ns. The expected error in the free energy calculations was the root-mean-square deviation in the energies of the sample in production period divided by the square root of the number of independent samples in the production period.[49]

$$\text{expected error} = \frac{\text{sample rms}}{\sqrt{\text{number of independent samples}}} \qquad (7)$$

## 3. Results

**3.1. Comparison between Predicted Binding Affinity and ITC Data.** *3.1.1. Calculations of Absolute Binding Free Energy by MM-PBSA and MM-GBSA Methods.* To evaluate the reproducibility and convergence of our free-energy calculation results, the same MM-GBSA protocol was applied to three independent 20 ns MD simulation runs of each of the WT−DRV, FLAP+−DRV, and ACT−DRV systems starting from each of their corresponding crystal structure (see the Methods section). To study the structural stability of the systems, the root-mean-square displacements (rmsd) of the Cα atoms of the simulated proteins were plotted over time with respect to their corresponding crystal structures (Supporting Information, Figure S2). For all the DRV−protease systems after 2 ns of MD simulations, the rmsd values were approximately 1.5 Å. As the calculations all require extensive equilibration, the averages of potential production periods were evaluated. After 10 ns simulations, the calculated binding free energy for DRV-WT stabilized (Figure 3A) for all three parallel simulations. Each of the triplicates of DRV−FLAP+ and DRV−ACT stabilized within 6 and 9 ns, respectively (Supporting Information, Figure S1). Thus, the first 10 ns was used as the equilibration period, as the free energy did not converge between the runs

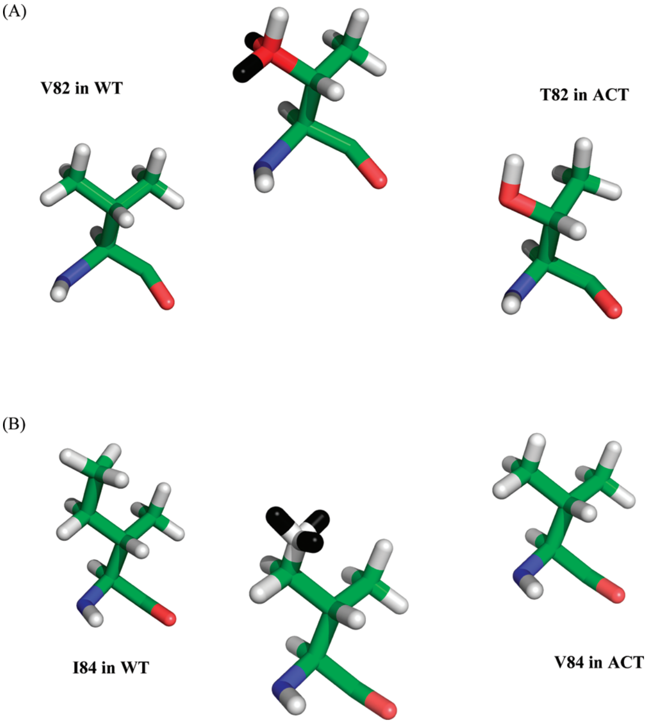


***Figure 2.*** Perturbation of Val82 to Thr and Ile84 to Val. Hydrogen atoms are colored white, oxygen atoms are colored red, nitrogen atoms are colored blue, carbon atoms are colored green, and dummy atoms are colored black. Left: residue in the wild-type protease as the initial state. Middle: the hybrid residue in the calculation process. Right: the mutated residue as end state. (A) The perturbation of Val82 to Thr82. (B) The perturbation of Ile84 to Val84.

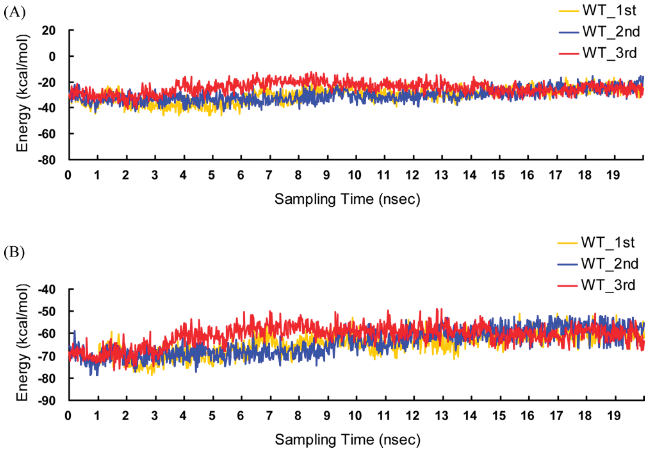


***Figure 3.*** (A) MM-GBSA calculated results of DRV−protease binding free energy with respect to the time. The three curves represent three independent MD trajectories. (B) vdW energy component of DRV−WT binding free energy with respect to the time.

***Table 1.*** Results of MM-GBSA Calculation for Absolute Binding Free Energy (kcal/mol) of DRV−Protease Based on Equilibration (1−10 ns) and Production (11−20 ns) Periods

| protease | sampling time (ns) | run 1 | run 2 | run 3 | average |
|---|---|---|---|---|---|
| WT | 1−10 | −33.4 | −32.8 | −24.9 | −30.4 |
|  | 11−20 | −27.3 | −27.0 | −25.7 | −26.7 |
| Flap+ | 1−10 | −23.4 | −17.7 | −20.1 | −20.4 |
|  | 11−20 | −21.1 | −20.8 | −21.1 | −21.0 |
| ACT | 1−10 | −20.3 | −13.9 | −25.5 | −19.9 |
|  | 11−20 | −17.8 | −17.6 | −19.8 | −18.4 |

(Table 1), while the second 10 ns was used as the production period, since generally the runs were converged.

The average predicted binding free energy of WT−DRV was −26.7 kcal/mol, that of FLAP+−DRV was −21.0 kcal/mol, and that of ACT−DRV was −18.4 kcal/mol (Table 2). Although these values differed from the ITC experimental values for each system (−15.2 kcal/mol for WT−DRV, −14.2 kcal/mol for FLAP+−DRV, and −13.6 kcal/mol for ACT−DRV), they correctly ranked the three protease variants' binding free energies: WT > FLAP+ > ACT. The more rigorous and time-consuming PB method was also used to calculate the polar solvation free energy. With this method, the predicted results were in better agreement with the ITC experimental data: −15.1 kcal/mol for WT−DRV, −11.6 kcal/mol for FLAP+−DRV, and −10.5 kcal/mol for ACT−DRV (Table 2). Comparison of the predicted polar solvation free energy difference calculated using the GB and PB models showed that the GB model had underestimated the polar solvation free energy of all three systems. This difference in estimates of polar solvation free energy by the GB and PB models has been reported and discussed in several studies involving different protein−ligand systems.[50−53] Such bias did not affect the ranking of binding energies for a given receptor with different ligands or for receptor variants with a specific ligand. Consistent with the results in other systems,[51] the MM-GBSA and MM-PBSA methods provide the same rank order of binding energies (Table 2) although the absolute values were different, for complexes of HIV-1 protease with DRV.

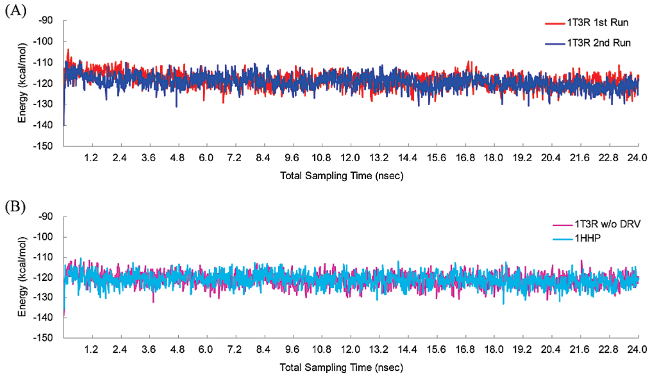


***Figure 4.*** Thermodynamic integration results over total sampling time showing that the calculations are stable. (A) $\Delta G_{\text{complex}}$ from two independent calculations of the coordinates of the DRV−WT crystal sturcture (1T3R). (B) $\Delta G_{\text{apo}}$ from two independent starting calculations from different starting structures: the DRV−WT crystal structure (1T3R) with the inhibitor removed is colored magenta, and the apo protease crystal structure (1HHP) is colored cyan.

*3.1.2. Calculation of Relative Binding Free Energy.* For the ACT double mutant V82T−I84V, the relative binding free energy was also calculated by the more rigorous and computationally more intensive thermodynamic integration method. This method has proven to be a powerful tool for studying binding free energy differences in a receptor−ligand system as statistical mechanics is its theoretical framework.[16,54] As described in the Methods section, thermodynamic integration calculated the binding free energy change from WT−DRV to ACT−DRV. The free energy changes associated with the alchemical pathways $\Delta G_{\text{apo}}$ and $\Delta G_{\text{complex}}$, which were the sum of 12 weighted $dV/d\lambda$ values (see eq 5, Table S1, Supporting Information), were plotted versus the time for the study of calculation convergence (Figure 4). For thermodynamic integration calculations, their reproducibility and internal consistency were studied by setting up two sets of independent simulations. Comparison of the two calculations of $\Delta G_{\text{complex}}$, which were started from the DRV−WT complex crystal structure coordinates, resulted in using the first 0.5 ns of each of the 12 $\lambda$ values as the equilibration period and the second 1.5 ns of each of the 12 $\lambda$ values as the production period. The total time for the equilibration period and production period were 6 and 18 ns, respectively. The two $\Delta G_{\text{complex}}$ values were −119.3 kcal/mol for run 1

***Table 2.*** Difference between the MM-GBSA and MM-PBSA Calculations (kcal/mol)

| protease | $\Delta G_{\text{SOLV-GB}}$ | $\Delta G_{\text{SOLV-PB}}$ | $\Delta G_{\text{CAL-GB}}$ | $\Delta G_{\text{CAL-PB}}$ | $\Delta G_{\text{EXP}}$[a] |
|---|---|---|---|---|---|
| WT | 52.8 ± 0.2 | 64.4 ± 0.4 | −26.7 ± 1.8 | −15.1 ± 1.8 | −15.2 ± 0.3 |
| FLAP+ | 53.9 ± 0.2 | 63.3 ± 0.3 | −21.0 ± 1.5 | −11.6 ± 1.5 | −14.2 ± 0.1 |
| ACT | 52.3 ± 0.2 | 60.2 ± 0.4 | −18.4 ± 1.7 | −10.5 ± 1.7 | −13.6 ± 0.2 |

[a] Experimental binding free energy data were obtained by ITC[5,66]

***Table 3.*** Thermodynamic Integration Calculation over 12 $\lambda$ on the Equilibration and Production Periods (kcal/mol)

| | $\Delta G_{\text{complex}}$ | | $\Delta G_{\text{apo}}$ | | |
|---|---|---|---|---|---|
| period | 1T3R first run | 1T3R second run | 1HHP | 1T3R w/o DRV | $\Delta\Delta G^{a}$ |
| equilibration period[b] | −116.1 ± 0.2 | −118.1 ± 0.2 | −120.7 ± 0.2 | −119.9 ± 0.2 | 3.2 ± 0.4 |
| production period[b] | −119.3 ± 0.1 | −119.9 ± 0.1 | −121.3 ± 0.1 | −121.5 ± 0.1 | 1.8 ± 0.2 |

[a] $\Delta\Delta G$ = Mean($\Delta G_{\text{complex}} - \Delta G_{\text{apo}}$). [b] Note that the equilibration period is the first 0.5 ns of each of the 12 $\lambda$'s, and the production period is the second 1.5 ns of each of the 12 $\lambda$'s, of the entire calculation. Total equilibration time is 6 ns, and production time is 18 ns.

***Table 4.*** Relative Binding Free Energy (kcal/mol) of ACT and WT HIV-1 Protease Calculated by Thermodynamic Integration, MM-GBSA, and MM-PBSA Methods vs ITC Data[5]

| | thermodynamic integration | MM-GBSA | MM-PBSA | ITC |
|---|---|---|---|---|
| $\Delta\Delta G$ | 1.8 ± 0.2 | 8.3 ± 3.5 | 4.6 ± 3.5 | 1.6 ± 0.5 |

and −119.9 kcal/mol for run 2 (Table 3). The $\Delta G_{\text{apo}}$ values calculated from the 1HHP and 1T3R crystal structure coordinates were −121.3 and −121.5 kcal/mol, respectively (Table 3). The protease in the 1HHP crystal structure has a flap semiopen conformation. The one in the 1T3R crystal structure has a flap close conformation. The free energy change ($\Delta G_{\text{Configurational}}$) associated with the transition from the semiopen conformations to the closed conformations protease was related to $\Delta G_{\text{apo}}$ as shown in the thermodynamic cycle below: in which

$$\begin{tikzcd}
\mathrm{WT}_{\text{semi-open}} \arrow[r, "\Delta G^{\text{WT}}_{\text{Configurational}}"] \arrow[d, "\Delta G^{\text{1HHP}}_{\text{apo}}"] & \mathrm{WT}_{\text{close}} \arrow[d, "\Delta G^{\text{1T3R}}_{\text{apo}}"] \\
\mathrm{ACT}_{\text{semi-open}} \arrow[r, "\Delta G^{\text{ACT}}_{\text{Configurational}}"] & \mathrm{ACT}_{\text{close}}
\end{tikzcd}$$

$$\Delta G^{\text{1HHP}}_{\text{apo}} - \Delta G^{\text{1T3R}}_{\text{apo}} = \Delta G^{\text{WT}}_{\text{Configurational}} - \Delta G^{\text{ACT}}_{\text{Configurational}}$$

The highly similar results of $\Delta G_{\text{apo}}$ calculated from both 1HHP and 1T3R structures indicated that the WT and the ACT had similar $\Delta G_{\text{Configurational}}$ values between their semiopen and close conformations. The relative binding free energy between DRV−WT and DRV−ACT was 1.8 kcal/mol. This result was a better match with the experimental relative binding free energy of 1.6 kcal/mol than 4.6 and 8.3 kcal/mol, which were calculated from the MM-PBSA and MM-GBSA methods, respectively (Table 4).

**3.2. Free-Energy Decomposition Analysis.** *3.2.1. Analysis of Contributions from Different Energy Components.* A free energy component analysis was performed to elucidate the mechanism for resistance to DRV of FLAP+ and ACT. The different energy components in the MM-PB/GBSA model (see the Methods section) were shown in Figure 5 and tabulated in the Supporting Information in more detail (Table S2). Both translational entropy ($-T\Delta S_{\text{translational}}$) change and rotational entropy change ($-T\Delta S_{\text{rotational}}$) were close in value in DRV binding for the three protease variants (Supporting Information, Table S2). They represented at least 90% of the change in entropy. The remaining vibration entropy change ($-T\Delta S_{\text{vibrational}}$) was 1.5 kcal/mol for DRV−WT binding, 2.2 kcal/mol for DRV−Flap+, and 0.2 kcal/mol for DRV−ACT. Further free-energy component analysis revealed that the favorable electrostatic interaction energy term ($\Delta G_{\text{ELE}}$) from the molecular mechanical energy ($\Delta G_{\text{MM}}$) had been canceled by the unfavorable polar solvation energy ($\Delta G_{\text{GB}}$) penalty. This result was in agreement with other MM-PB/GBSA studies.[24,55−58] The total electrostatic interaction energy ($\Delta G_{\text{ELE}} + \Delta G_{\text{GB}}$) for DRV−WT was 15.4 kcal/mol, for DRV−FLAP+ was 14.1 kcal/mol, and for DRV−ACT was 17.0 kcal/mol. The vdW interaction energy was −60.3 kcal/mol for DRV−WT, −54.5 kcal/mol for DRV−FLAP+, and −52.8 kcal/mol for DRV−ACT. The vdW interactions had the largest contribution to protease−inhibitor binding (Figure 5A) and sustained the largest energy loss in both the FLAP+ and ACT drug-resistant mutants (Figure 5B).

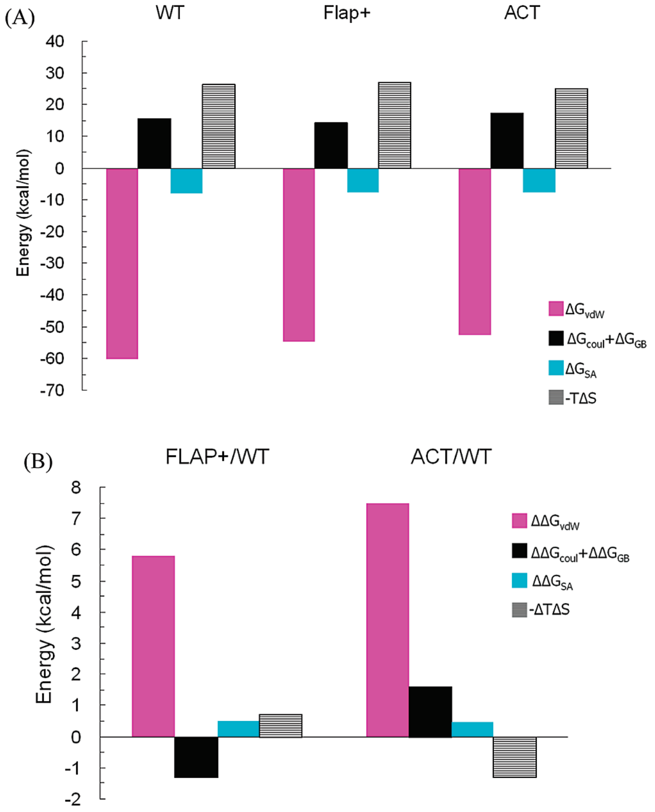


***Figure 5.*** (A) Binding free energy components of DRV−WT, DRV−FLAP+, and DRV−ACT. (B) The loss of binding free energy components with DRV of FLAP+ and ACT compared to the WT protease.

*3.2.2. Free Energy Projected to Each Residue of HIV-1 Protease.* In order to gain extra insight into the mechanisms of protease−inhibitor binding and drug resistance, the binding free energy calculated from the MM-GBSA method had been broken down to individual protease residues. The

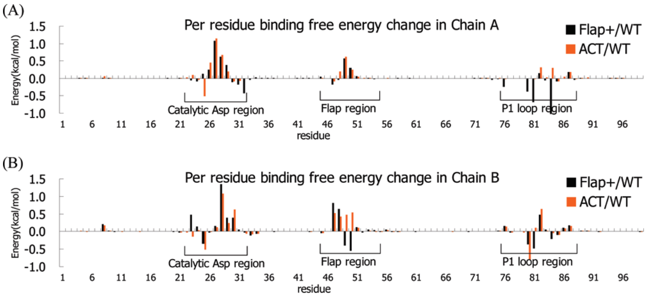


***Figure 6.*** Decomposition of energy from MM-GBSA per residue of HIV-1 protease. (A) Energy difference between wild-type protease and FLAP+ variant. (B) Energy difference between wild-type protease and ACT variant.

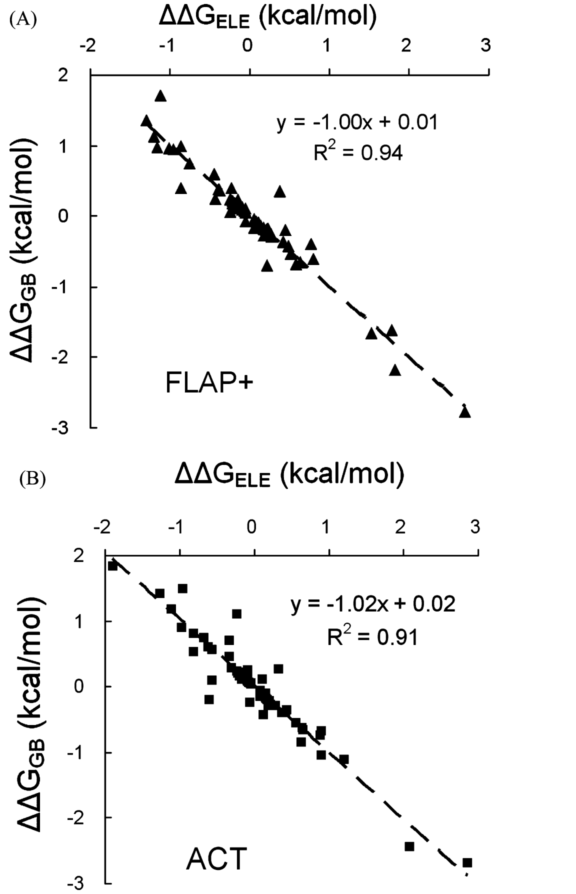


***Figure 7.*** Correlation between $\Delta\Delta G_{\mathrm{ELE}}$ and $\Delta\Delta G_{\mathrm{GB}}$ of each residue. (A) Energy difference between FLAP+ and WT. (B) Energy difference between ACT and WT.

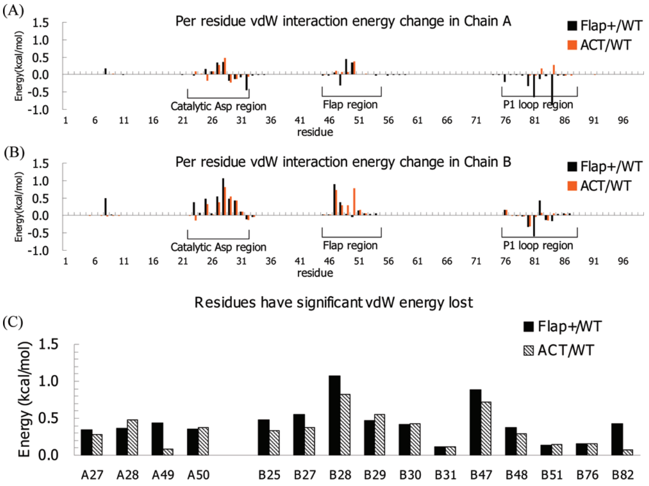


***Figure 8.*** (A) vdW energy loss between FLAP+ and WT protease. (B) vdW energy loss between ACT and WT protease. (C) Residues with a vdW energy loss larger than 0.1 kcal/mol.

energy difference was investigated between the WT−DRV complex and the two drug-resistant mutant protease−DRV complexes for each residue (Figure 6). The residues with energy changes were mainly located in three areas (Supporting Information, Figure S3), the catalytic region (residues 22 to 33), the flap region (residues 45 to 55), and the P1 loop region (residues 79 to 87) on both monomers of the protease. These energy changes varied asymmetrically in the two protease monomers. Many residues (27, 28, 50, 87, 8′, 28′, 29′, 47′, and 76′) structurally adjacent to DRV other than those that mutate (10, 48, 54, and 82 for FLAP+; 82 and 84 for ACT) responded to the mutations (Figure 6) as had been previously observed.[14,28] The sites of mutation not only impacted their own binding free energy interactions with inhibitors but also influenced the interaction of other residues with the inhibitor by inducing alterations in the geometry of the binding site.

Favorable electrostatic interactions opposed by the polar solvation energy penalty also apply to individual residues. A change in electrostatic energy ($\Delta\Delta G_{\mathrm{ELE}}$) of any residue was always associated with an equal but opposite compensation in solvation energy ($\Delta\Delta G_{\mathrm{GB}}$) of very similar amplitude but in a different direction. The correlation coefficient for the $\Delta\Delta G_{\mathrm{ELE}}$ and $\Delta\Delta G_{\mathrm{GB}}$ of FLAP+ was −0.97 and for the ACT was −0.95 (Figure 7A for FLAP+, Figure 7B for ACT). This high correlation of $\Delta\Delta G_{\mathrm{ELE}}$ and $\Delta\Delta G_{\mathrm{GB}}$ made the change of vdW energy the largest factor in the loss of binding free energy between DRV and FLAP+/ACT. The residues in the catalytic, flap, and P1 loop regions also had the largest change in vdW interaction energy (Figure 8). To highlight those residues with a significant difference between the WT and the two drug-resistant mutants, a cutoff of 0.1 kcal/mol of vdW energy change was used. The residues in FLAP+ and ACT with a loss of vdW energy greater than the cutoff were plotted in Figure 8C. In chain A, these residues included 26, 27, 28, 47, 49, and 50; in chain B, these residues were 8′, 25′, 27′−31′, 47′−49′, 51′, 52′, 54′, 76′, 82′, and 86′ (Supporting Information Figure S3D). The loss in vdW interaction energy of chain B was significantly larger than that of chain A.

*3.2.3. vdW Energy Contribution from Each DRV Atom.* To explore the mechanism of the loss in binding free energy between DRV and the drug-resistant mutants, the vdW energy contributions were calculated for each DRV atom and compared between complexes with the WT and FLAP+ and ACT mutant proteases. DRV had 75 atoms, of which 37 were hydrogen atoms with very limited contribution to the vdW interaction energy. Thus, data were presented for only the 38 heavy atoms in DRV (Figure 9A). Structurally, DRV could be considered formed by four major moieties:

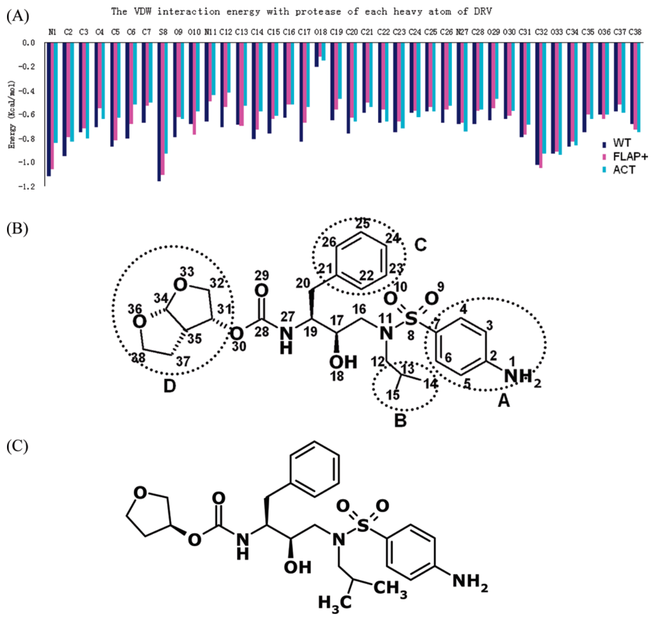


*Figure 9.* (A) vdW interaction energy of each non-hydrogen atom of DRV with protease. The energy of DRV−WT is colored dark blue, the energy of DRV−FLAP+ is colored magenta, and the energy of DRV−ACT is colored cyan. (B) The definition of four moieties of DRV. (C) The chemical structure of APV.

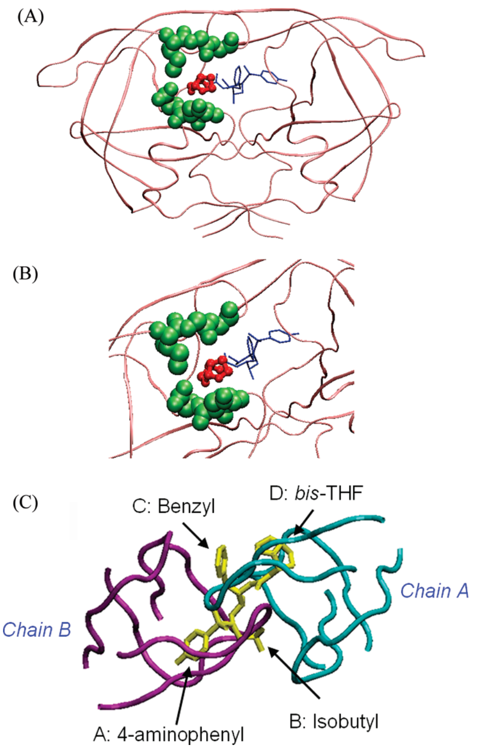


*Figure 10.* (A and B) Cluster of vdW contacts formed by the *bis*-THF group and the protease residues Ala28, Asp29, Asp30, Ile47, Gly48, and Gly49 of chain A. The atoms of the above residues are displayed and colored green. The atoms of the bis-THF group are colored red, while the rest of DRV is colored blue. (C) Relative position of DRV's four moieties (colored yellow) to chain A (colored cyan) and chain B (colored purple) of protease.

(A) the 4-amino phenyl group, (B) the isopropyl group, (C) the benzyl ring, and (D) *bis*-tetrahydrofuranylurethane (THF; Figure 9B). To compare the energy change between these 4 moieties, we defined DV (loss of vdW interaction energy ratio) as

$$\mathrm{DV} = \frac{\sum_i (E_i^{\mathrm{Flap+}} - E_i^{\mathrm{WT}}) + \sum_i (E_i^{\mathrm{ACT}} - E_i^{\mathrm{WT}})}{2 \times \sum_i E_i^{\mathrm{WT}}} \times 100\%$$

where *i* is the atom within a specific moiety. The *bis*-THF moiety and the benzyl ring have relatively low DVs of 3.1% and 8.5%, respectively. The 4-amino phenyl and isobutyl groups have significantly higher DVs of 17.0% and 19.2%, respectively (Table 5).

The major difference between DRV and a previous generation protease inhibitor, amprenavir (APV; Figure 9C), is that DRV has a second tetrahydrofuran ring, which is part of its *bis*-THF moiety. Nonetheless, DRV has been shown by ITC experiments[5] to bind more tightly than APV with the protease, with a 2.6 kcal/mol larger binding affinity. In the DRV−WT protease structure (1T3R), the *bis*-THF moiety is surrounded by the protease chain A residues Ala28, Asp29, Asp30, Ile47, Gly48, and Gly49, which form a cluster of vdW contacts (Figure 10A and B). This packing can be also observed from the crystal structures of the DRV−FLAP+ and DRV−ACT complexes.[5] Examination of the MD simulation structures of DRV in complex with the WT, FLAP+, and ACT proteases showed that these residues and the *bis*-THF moiety maintained a relatively stable structure compared to other parts of the inhibitor. This stability led to the small ratio of the *bis*-THF group's vdW energy loss (Table 5).

Similar to the *bis*-THF group, the benzyl ring maintained its vdW interactions with protease residues in chain A (Figure 10C) in most conformations sampled by the MD simulations. This stability in DRV interactions with chain A explained the asymmetric vdW energy losses of the protease's two chains. Chain B was shown by free-energy decomposition of protease residues to have more residues with significant energy loss than chain A (Figure 8C). Unlike the *bis*-THF group and the benzyl ring, whose vdW interactions were only slightly influenced by the drug-resistant mutations, the 4-amino phenyl and isobutyl groups of DRV in complex with FLAP+ and ACT lost approximately 20% of the vdW interaction energy. A comparison of the MD simulation structure of DRV−WT with those of the two drug-resistant mutants showed that the 4-amino phenyl and isobutyl groups of DRV in the DRV−FLAP+ and DRV−ACT complexes

*Table 5.* Loss of van der Waals' Interaction Energy (DV) for Different DRV Moieties

| | 4-amino phenyl group | isobutyl group | benzyl ring | *bis*- tetrahydrofuranyl |
|---|---|---|---|---|
| DV (%) | 17 | 19 | 9 | 3 |
| number of heavy atoms | 7 | 4 | 7 | 8 |

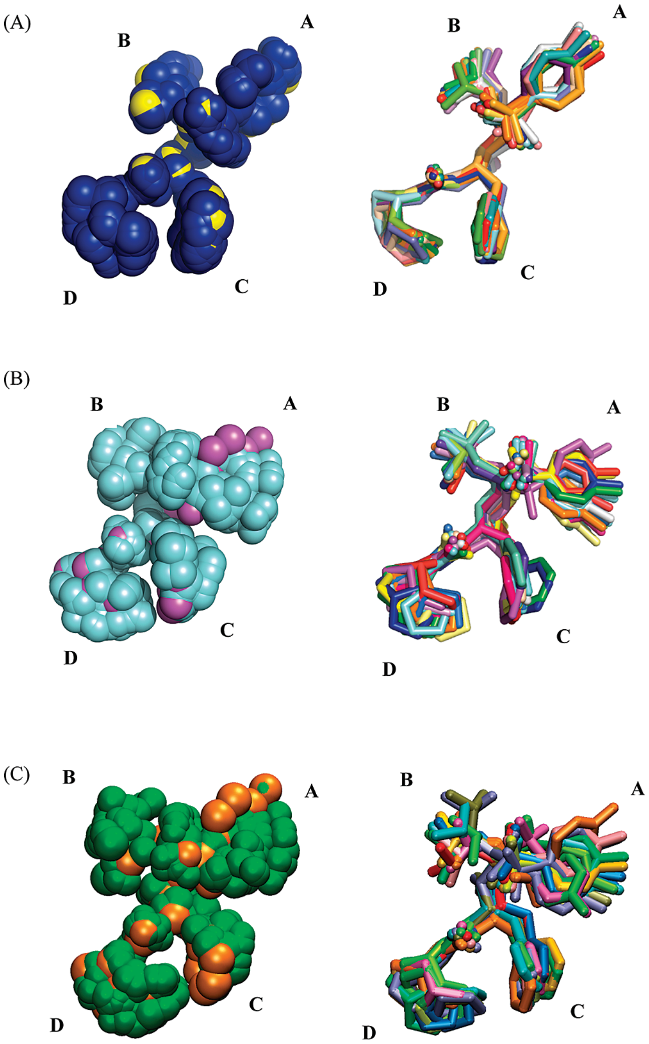


***Figure 11.*** (A) Conformational space of DRV sampled in DRV−WT complex simulations. Left: DRV ensemble is shown with atoms' vdW radii. The original conformation as in the crystal structure is colored yellow. The sampled conformations ensemble from MD simulation is colored blue. Right: 20 snapshots of DRV conformations taken every 1 ns from MD simulations. (B) Conformational space of DRV sampled in DRV−FLAP+ complex simulations. Left: DRV ensemble is shown with atoms' vdW radii. The original conformation as in the crystal structure is colored purple. The sampled conformations ensemble from MD simulation is colored cyan. Right: 20 snapshots of DRV conformations taken every 1 ns from MD simulations. (C) Conformational space of DRV sampled in DRV−ACT complex simulations. Left: DRV ensemble is shown with atoms' vdW radii. The original conformation as in the crystal structure is colored orange. The sampled conformations ensemble from MD simulation is colored green. Right: 20 snapshots of DRV conformations taken every 1 ns from MD simulations.

undergo significant geometry changes (Figure 11A and B) that led to these groups losing their vdW contacts with the drug-resistant proteases.

The free-energy decomposition by residue showed that the mutations had induced changes in the shape of the binding pocket as evidenced by the predominant changes occurring in the vdW interactions energy. Overall, there was a decrease in the vdW interaction energy between the protease and DRV, mostly on the 4-amino phenyl side, as the volume of the binding pocket was effectively enlarged as the mutations within the active site were to smaller residues (V82A in Flap+ and I84V in ACT). This expansion of the active site permits, as we had observed, other residues to interact to varying degrees with the inhibitor; in this way, the FLAP+ and ACT mutant proteases could develop drug resistance.

## 4. Conclusion and Discussion

With the appearance of drug-resistant HIV-1 protease variants becoming one of the major challenges to AIDS therapy, understanding the mechanism of drug resistance is critical. This goal is best addressed by cross-analyzing the data on protease mutants from different experimental methods such as crystallography and isothermal titration calorimetry.[59–62] Comparing the crystal structure of APV bound to wild-type protease and a drug-resistant protease variant, King et al.[5] found that the mutation I84V has decreased the vdW interaction between APV and the drug-resistant variant, which might account for the loss of binding affinity between APV and the drug-resistant variant. By analyzing the ITC experiments results, Luque et al.[59] suggested that the drug-resistant mutations change the shape of the active site. The very flexible substrates are less susceptible to the change than the synthetic inhibitors,[59] which might enable the drug-resistant protease variant to still recognize the substrate while having less binding affinity with the synthetic inhibitors. Comparing the trajectories from MD simulations on the wild-type protease and the V82F/I84V protease variant, Perryman et al.[63] suggested that the mutations changing the equilibrium between the flap-semiopen and closed conformations could be one aspect of the protease drug-resistant mechanism. More details about inhibitor−protease binding can be provided by free energy calculations, which start from structural coordinates and yield thermodynamic data. In this study, we performed MM-PB/GBSA calculations and free-energy component analysis of DRV−WT, DRV−FLAP+ (L10I, G48V, I54V, V82A), and DRV-ACT (V82T, I84V). By running three independent 20 ns simulations for each of these systems, we not only identified the convergence and consistency of our calculations but also predicted the order of binding energies in agreement with ITC data. As described in the Methods section, the calculations of solvation energy and molecular mechanic energy were based on 1000 frames with a 20 ps interval. The more time-consuming entropy calculation was based on 100 frames with a 200 ps interval. In order to examine the statistical significance of the binding free energy of protease with DRV, we calculated the entropy for each frame that was used to calculate the solvation energy and the molecular mechanic energy. The difference of calculated entropy using 100 frames and 1000 frames was tabulated in the Supporting Information (Table S3). $t$ tests were performed to evaluate the significance of the difference of $\Delta G_{\text{MM-PBSA}}$ and $\Delta G_{\text{MM-GBSA}}$ between WT, Flap+, and ACT. The $p$ values were all less than 0.01, which indicated significant differences.

Moreover, the relative binding free energy between DRV−WT and DRV−ACT using MM-PB/GBSA and thermodynamic integration (TI) methods was calculated. The accuracy of these result had the rank order TI > MM-PBSA > MM-GBSA, which is the same order of the computational times required for these methods. The TI method is more suitable for comparing the energy difference between two similar systems. In the case of the ACT (V82T, I84V) mutant here, the TI method not only gave the more accurate predicted energy than the MM-PB/GBSA method but also had better reproducibility and faster convergence (Figure 3, Figure 4).

The results of free energy components analysis showed that the vdW interaction energy was dominant in the total binding free energy change. The contribution from charged interactions was minor compared to vdW interactions due to the cancellation of electrostatic interactions energy and the polar desolvation energy. Interestingly, a previous free energy component analysis showed that, compared to the total predicted binding free energy, the predicted contribution from electrostatic interaction had a higher correlation with the experimental binding free energy.[64] Recently, a similar analysis from the same group on interactions between HIV protease and inhibitors concluded that the total theoretical binding energy was in agreement with the experimental data, although the free energy component from only the charged interactions was also correlated well with experimental binding free energy.[65] This difference might have resulted from the different environments of the two systems.[65] The former calculation was on the large solvated protein surface, while the latter calculation on HIV protease was on a relatively small and buried binding pocket. The free-energy decomposition analysis on protease residues indicated that mutations in the protease induced conformational changes in its active site. The *bis*-THF group and benzyl ring of DRV sustained their vdW interactions with the drug-resistant protease variants and contribute most to the inhibitor−protease binding, while DRV's 4-amino phenyl and isobutyl groups were susceptible to changes in the protease's binding pocket and adopted conformations that lose vdW interaction with drug-resistant variants (Table 5).

These findings suggested that the design of new protease inhibitors based on the DRV scaffold should consider reoptimizing 4-aminophenyl and isopropyl groups since these parts of DRV did not maintain their interactions with drug resistant protease variants as much as the *bis*-THF group. Such new inhibitors would likely bind more tightly to HIV protease and may be less susceptible to drug resistance.

## Abbreviations

Bis-THF, bis-tetrahydrofuranyl; ITC, isothermal titration calorimetry; TI, thermodynamic integration; MM-PBSA, molecular mechanics−Poisson−Boltzmann surface area; GB, generalized Born; DRV, Darunavir; ACT, HIV-1 protease variant V82T, I84 V; FLAP+, HIV-1 protease variant L10I, G48V, I54 V, V82A; vdW, van der Waals; MD, molecular dynamics; ns, nanosecond; ps, picosecond; fs, femtosecond.

**Acknowledgment.** We acknowledge Nancy King, Moses Prabu, Ellen Nalivaika, and Madhavi Nalam in the acquisition of the experimental data. We also acknowledge Claire Baldwin for the preparation of the manuscript. This research was supported by the National Institutes of Health (NIH), grant P01-GM66524. Calculations were performed on a computer cluster purchased with NIH 1S10440166892993.

**Supporting Information Available:** Figures of MM-GBSA calculated results of DRV-Flap+, DRV-ACT binding free energy with respect to the time; plots of the rmsd of Cα atoms of protease with respect to their corresponding crystal structures over time; Tables S1 and S2. This material is available free of charge via the Internet at http://pubs.acs.org/.

## References

(1) Debouck, C. *AIDS Res. Hum. Retroviruses* **1992**, *8*, 153.

(2) Wlodawer, A.; Erickson, J. W. *Annu. Rev. Biochem.* **1993**, *62*, 543.

(3) Wood, E.; Hogg, R. S.; Yip, B.; Moore, D.; Harrigan, P. R.; Montaner, J. S. *HIV Med.* **2007**, *8*, 80.

(4) Schinazi, R. F.; Larder, B. A.; Mellors, J. W. *Int. Antiviral News* **1997**, *5*, 129.

(5) King, N. M.; Prabu-Jeyabalan, M.; Nalivaika, E. A.; Wigerinck, P.; de Bethune, M. P.; Schiffer, C. A. *J. Virol.* **2004**, *78*, 12012.

(6) Todd, M. J.; Luque, I.; Velazquez-Campoy, A.; Freire, E. *Biochemistry* **2000**, *39*, 11876.

(7) King, N. M.; Melnick, L.; Prabu-Jeyabalan, M.; Nalivaika, E. A.; Yang, S. S.; Gao, Y.; Nie, X.; Zepp, C.; Heefner, D. L.; Schiffer, C. A. *Protein Sci.* **2002**, *11*, 418.

(8) Prabu-Jeyabalan, M.; Nalivaika, E. A.; King, N. M.; Schiffer, C. A. *J. Virol.* **2003**, *77*, 1306.

(9) Talhout, R.; Villa, A.; Mark, A. E.; Engberts, J. B. *J. Am. Chem. Soc.* **2003**, *125*, 10570.

(10) Huang, N.; Jacobson, M. P. *Curr. Opin. Drug Discovery Dev.* **2007**, *10*, 325.

(11) Jorgensen, W. L. *Science* **2004**, *303*, 1813.

(12) Bash, P. A.; Singh, U. C.; Brown, F. K.; Langridge, R.; Kollman, P. A. *Science* **1987**, *235*, 574.

(13) Wang, W.; Kollman, P. A. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 14937.

(14) Wittayanarakul, K.; Aruksakunwong, O.; Sompornpisut, P.; Sanghiran-Lee, V.; Parasuk, V.; Pinitglang, S.; Hannongbua, S. *J. Chem. Inf. Model.* **2005**, *45*, 300.

(15) Adcock, S. A.; McCammon, J. A. *Chem. Rev.* **2006**, *106*, 1589.

(16) Gao, J.; Kuczera, K.; Tidor, B.; Karplus, M. *Science* **1989**, *244*, 1069.

(17) Zwanzig, R. W. *J. Chem. Phys.* **1954**, *22*, 1420.

(18) Michielin, O.; Karplus, M. *J. Mol. Biol.* **2002**, *324*, 547.

(19) Archontis, G.; Simonson, T.; Moras, D.; Karplus, M. *J. Mol. Biol.* **1998**, *275*, 823.

(20) Singh, U. C.; Benkovic, S. J. *Proc. Natl. Acad. Sci. U. S. A.* **1988**, *85*, 9519.

(21) Lawrenz, M.; Baron, R.; McCammon, J. A. *J. Chem. Theory Comput.* **2009**, *5*, 1106.

(22) Massova, I.; Kollman, P. A. *Perspect. Drug Discovery Des.* **1999**, *18*, 113.

(23) Xu, Y.; Wang, R. *Proteins* **2006**, *64*, 1058.

(24) Gohlke, H.; Kiel, C.; Case, D. A. *J. Mol. Biol.* **2003**, *330*, 891.

(25) Archontis, G.; Simonson, T.; Karplus, M. *J. Mol. Biol.* **2001**, *306*, 307.

(26) Swanson, J. M.; Henchman, R. H.; McCammon, J. A. *Biophys. J.* **2004**, *86*, 67.

(27) Hendsch, Z. S.; Tidor, B. *Protein Sci.* **1999**, *8*, 1381.

(28) Hou, T.; Yu, R. *J. Med. Chem.* **2007**, *50*, 1177.

(29) Surleraux, D. L.; Tahri, A.; Verschueren, W. G.; Pille, G. M.; de Kock, H. A.; Jonckers, T. H.; Peeters, A.; De Meyer, S.; Azijn, H.; Pauwels, R.; de Bethune, M. P.; King, N. M.; Prabu-Jeyabalan, M.; Schiffer, C. A.; Wigerinck, P. B. *J. Med. Chem.* **2005**, *48*, 1813.

(30) Surleraux, D. L.; de Kock, H. A.; Verschueren, W. G.; Pille, G. M.; Maes, L. J.; Peeters, A.; Vendeville, S.; De Meyer, S.; Azijn, H.; Pauwels, R.; de Bethune, M. P.; King, N. M.; Prabu-Jeyabalan, M.; Schiffer, C. A.; Wigerinck, P. B. *J. Med. Chem.* **2005**, *48*, 1965.

(31) Case, D. A.; Cheatham, T. E., III; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668.

(32) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999.

(33) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157.

(34) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; 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.; Gonszalez, C.; Pople, J. A. *Gaussian 03*, Revision B.05; Gaussian, Inc.: Pittsburgh, PA, 2003.

(35) Bayly, C. I. C., P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269.

(36) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327.

(37) Onufriev, A.; Case, D. A.; Bashford, D. *J. Comput. Chem.* **2002**, *23*, 1297.

(38) Onufriev, A.; Bashford, D.; Case, D. A. *Proteins* **2004**, *55*, 383.

(39) Luo, R.; David, L.; Gilson, M. K. *J. Comput. Chem.* **2002**, *23*, 1244.

(40) Sitkoff, D.; Sharp, K. A.; Honig, B. *J. Phys. Chem.* **1994**, *98*, 1978.

(41) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. *Biophys. J.* **1997**, *72*, 1047.

(42) Kollman, P. A. *Chem. Rev.* **1993**, *93*, 2395.

(43) Radmer, R. J.; Kollman, P. A. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 215.

(44) Lau, F. T.; Karplus, M. *J. Mol. Biol.* **1994**, *236*, 1049.

(45) Blondel, A. *J. Comput. Chem.* **2004**, *25*, 985.

(46) Wong, C. F.; McCammon, J. A. *J. Am. Chem. Soc.* **1986**, *108*, 3830.

(47) Case, D. A.; Cheatham, T. E., III; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668.

(48) Lee, T.; Kollman, P. A. *J. Am. Chem. Soc.* **2000**, *122*, 4385.

(49) Bishop, M.; Frinks, S. *J. Phys. Chem.* **1987**, *87*, 3675.

(50) Bea, I.; Gotsev, M. G.; Ivanov, P. M.; Jaime, C.; Kollman, P. A. *J. Org. Chem.* **2006**, *71*, 2056.

(51) Wittayanarakul, K.; Hannongbua, S.; Feig, M. *J. Comput. Chem.* **2008**, *29*, 673.

(52) Gohlke, H.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 238.

(53) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L. *J. Med. Chem.* **2004**, *47*, 3032.

(54) Lee, T.; Kollman, P. A. *J. Am. Chem. Soc.* **2000**, *122*, 4385.

(55) Stoica, I.; Sadiq, S. K.; Coveney, P. V. *J. Am. Chem. Soc.* **2008**, *130*, 2639.

(56) Massova, I.; Kollman, P. A. *J. Am. Chem. Soc.* **1999**, *11*, 8133.

(57) Wang, W.; Kollman, P. A. *J. Mol. Biol.* **2000**, *303*, 567.

(58) Huo, S.; Massova, I.; Kollman, P. A. *J. Comput. Chem.* **2002**, *23*, 15.

(59) Luque, I.; Todd, M. J.; Gomez, J.; Semo, N.; Freire, E. *Biochemistry* **1998**, *37*, 5791.

(60) Todd, M. J.; Luque, I.; Velazquez-Campoy, A.; Freire, E. *Biochemistry* **2000**, *39*, 11876.

(61) Velazquez-Campoy, A.; Kiso, Y.; Freire, E. *Arch. Biochem. Biophys.* **2001**, *390*, 169.

(62) Ohtaka, H.; Velazquez-Campoy, A.; Xie, D.; Freire, E. *Protein Sci.* **2002**, *11*, 1908.

(63) Perryman, A. L.; Lin, J. H.; McCammon, J. A. *Protein Sci.* **2004**, *13*, 1108.

(64) Lippow, S. M.; Wittrup, K. D.; Tidor, B. *Nat. Biotechnol.* **2007**, *25*, 1171.

(65) Huggins, D. J.; Altman, M. D.; Tidor, B. *Proteins* **2009**, *75*, 168.

(66) King, N. M.; Prabu-Jeyabalan, M.; Bandaranayake, R. M.; Nalam, M. N.; Ozen, A.; Haliloglu, T.; Schiffer, C. In preparation 2009.

CT9004678

JCTC Journal of Chemical Theory and Computation

# Low Inhibiting Power of N···CO Based Peptidomimetic Compounds against HIV-1 Protease: Insights from a QM/MM Study

Julian Garrec,[†,||] Michele Cascella,[‡] Ursula Rothlisberger,[§] and Paul Fleurat-Lessard*[,†]

*Université de Lyon, École Normale Supérieure de Lyon, Laboratoire de Chimie − UMR 5182, 46 allée d'Italie, 69364 Lyon Cedex 07, France, Department of Chemistry and Biochemistry, University of Bern, Freiestrasse 3, CH-3012 Bern, Switzerland, and Laboratory of Computational Chemistry and Biochemistry, Institute of Chemical Sciences and Engineering, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland*

Received September 8, 2009

**Abstract:** Recently, Hasserodt et al. proposed new HIV-1 drug candidates based on a weak N···CO interaction, designed to be a close transition state analog (Gautier et al. *Bioorg. Med. Chem.* **2006**, *14*, 3835−3847; Waibel et al. *J. Bioorg. Med. Chem.* **2009**, *17*, 3671−3679). They suggested that further improvement of these compounds could take advantage of computational approaches. In the present work, we propose an atomistic model based on a QM/MM description of the N···CO core embedded in an amino-aldehyde peptidic inhibitor. We focus on the existence of the N···CO interaction in the aqueous and enzymatic media. We show that the N···CO bond holds in water, while in the protein, there is a competition between the formation of the weak N···CO bond and the conservation of the hydrogen bond network around the structural water molecule W301 that is known to be crucial for the binding of both substrates and inhibitors. This competition hampers the inhibitor to provide strong stabilizing interactions with all the key parts of the protein at the same time. Our calculations indicate that this competition we observed in peptidic compounds might be avoided by the proper design of nonpeptidic ones, following a similar strategy to that for cyclic urea derivatives and the FDA approved drug Tipranavir. Hence, our results encourage further development of the nonpeptidic hydrazino-urea derivatives suggested recently by Hasserodt et al.

## 1. Introduction

The human immunodeficiency virus type 1 aspartic protease (HIV-1 PR, Figure 1) is one of the major targets for the design of anti-AIDS drugs.[1,2] This enzyme catalyzes the hydrolysis at specific sites of the polyprotein encoded by the virus genome yielding separate functional proteins.[3−5] This function was shown early to be crucial to virion assembly and maturation, and its disruption by either active-site mutation or inhibition leads to the production of viral particles that lack infectious ability.[6−8]

Several HIV-1 PR inhibitors have been approved by the FDA and significantly prolong the life expectancy of HIV infected patients.[1,9−11] Nevertheless, the rapid emergence of resistance caused by multiple HIV-1 PR mutations decreases the effectiveness of these drugs.[9,12,13] Almost all FDA-approved drugs are peptidomimetic active-site inhibitors that contain a hydroxyl group designed to interact with the central

* To whom correspondence should be addressed. Phone: +33 4 7272 8154, Fax: +33 4 7272 8860. E-mail: Paul.Fleurat-Lessard@ens-lyon.fr.

[†] École Normale Supérieure de Lyon.

[‡] University of Bern.

[§] École Polytechnique Fédérale de Lausanne.

[||] Present address: Laboratory of Computational Chemistry and Biochemistry, Institute of Chemical Sciences and Engineering, École Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland.

10.1021/ct9004728 © 2010 American Chemical Society
Published on Web 03/26/2010

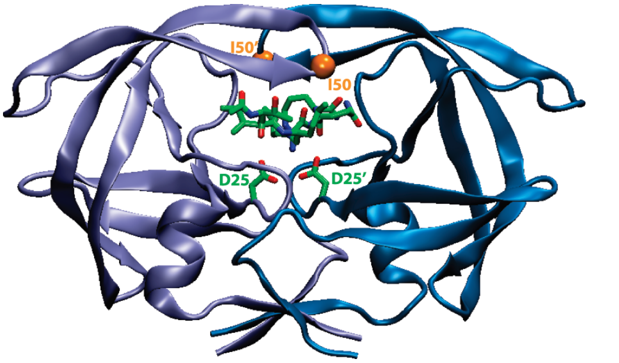


***Figure 1.*** Crystallographic structure of HIV-1 PR in complex with the peptidomimetic inhibitor MVT-101[36,34] (4HVP entry in the PDB data bank[44]). The *aspartyl dyad* (residues 25 and 25′) is located in the lower part of the active site. Ile 50 and 50′ are located at the flap tips and are indicated with orange spheres. Drawings were made with the VMD program.[83]

aspartyl dyad of HIV-1 PR (see the Abbreviations section for a list of abbreviations used in this work).[1,2]

Freire et al. suggested that the competitive advantage of HIV-1 PR inhibitors over substrate binding is probably due to their higher rigidity, providing a more favorable entropic change upon binding. However, because of their rigidity, these inhibitors are less amenable to adapt to shape modifications of the enzymatic binding site induced by mutations.[14] Hence, the design of new potent inhibitors that exhibit both a better binding free energy and an increased flexibility remains a challenging task.[13,15−17]

In that context, finding functional groups that yield stronger interactions with the aspartyl dyad than the usual hydroxyl group would be of great interest.[18] Toward this aim, Hasserodt et al.[19−21] proposed a new concept of aspartic protease inhibitors based on a noncovalent interaction of a tertiary amine nitrogen with a carbonyl group, the so-called N···CO bond, that they believed to be a better transition state analog than the commonly used hydroxyl moiety (Figure 2). As a first attempt, they synthesized a series of dipeptide mimics containing the N···CO core,[19] the amine and aldehyde fragments being bridged by an ethylene moiety (referred to as the *aliphatic bridge* hereafter, Figure 2a).

The best candidate of these amino-aldehyde peptides (AAP) exhibited an inhibition constant of 97 $\mu$M, far from the typical picomolar value that is desirable for a potent inhibitor.[15,17,22−25]

In order to design more efficient inhibitors, they synthesized new compounds based on hydrazyno-urea heterocycles containing both the N···CO core and a carbonyl group aimed at forming hydrogen bonds with two NH groups of the upper part of the HIV-1 PR active site,[20,21] the so-called *flaps*. This strategy is similar to that involved in the design of cyclic urea derivatives[26] and the FDA-approved drug Tipranavir.[27] The hydrazyno-urea derivatives indeed proved to interact slightly stronger with HIV-1 PR ($K_i \approx 29$ $\mu$M for the best candidate). Hasserodt et al. concluded that a hit to lead optimization could now be initiated with the help of computational methods.[20,21] Such an approach, however, needs the definition of a proper model describing the N···CO core in a biological environment. The purpose of the present work is to design this model and to apply it to AAPs in order to help elucidate the origin of their low inhibition power.

Using a mixed quantum mechanics/molecular mechanics (QM/MM)[28] approach, we have designed a model that explicitly includes the solvated enzyme complexed with an AAP. An accurate quantum level of theory is crucial to the correct description of the N···CO core interacting with the aspartyl dyad, due to the intrinsic complexity of this system. No accurate transferable force field parameters exist for the N···CO bond,[29,30] in particular because the stability of the N···CO is highly sensitive to the nature of the surrounding medium.[19,31−33] For instance, the N···CO bond is unstable in apolar-aprotic media, leaving the tertiary amine and the aldehyde groups essentially independent. The structure of a complex between HIV-1 PR and an AAP has not been reported yet. Our study aims at exploring the feasibility of N···CO bond formation in this enzyme. In order to compare the behavior of the AAP, in particular the N···CO bond, in the protein and in aqueous media, we have also designed a similar QM/MM model of the AAP in water.

This article is organized as follows: in the first part, we detail our structural model and the computational procedure used to model the AAP in the enzymatic and aqueous media. Results are reported and discussed in the second part, while the last part summarizes our main findings.

## 2. Materials and Methods

**2.1. Initial Structure.** No experimental structure of an AAP complexed with HIV-1 PR (E•AAP) is available. Nonetheless, on the basis of known structural features of HIV-1 PR−substrate complexes and previous kinetics/inhibition experiments of Hasserodt et al., it is possible to construct a starting structure for molecular dynamics simulations. Indeed, inhibition profiles show a competitive mechanism strongly suggesting that the AAP truly binds to the active site of the enzyme.[19] In addition, HIV-1 PR is known to bind a variety of peptide substrates in the same extended conformation,[34,35] the substrate backbone exhibiting many hydrogen bonds with the enzyme and the side chains being accommodated in a series of binding site subpockets. Due to the very high peptidic character of AAPs, their backbone and side chains should bind the active site in a way very similar to that of the corresponding peptides. Since AAPs are designed to be transition state analogs,[19] the reaction intermediate (E•INT) that connects the two TSs along the reaction pathway of HIV-1 PR appears as a natural starting structure for E•AAP modeling.

From both $^{18}$O isotope exchange experiments[36] and X-ray structures that captured key intermediate stages of the catalytic reaction,[37,38] there is evidence that the substrate peptide bond cleavage involves the nucleophilic attack of a water molecule onto the scissile peptide bond, leading to a tetrahedral intermediate (Figure 2c). Note that, despite this commonly accepted picture, the protonation state of the intermediate is still a matter of controversy. Indeed, while ab initio calculations suggested that the intermediate is a neutral gem-diol,[39,40] an empirical valence bond (EVB) model—calibrated against DFT calculations involving model

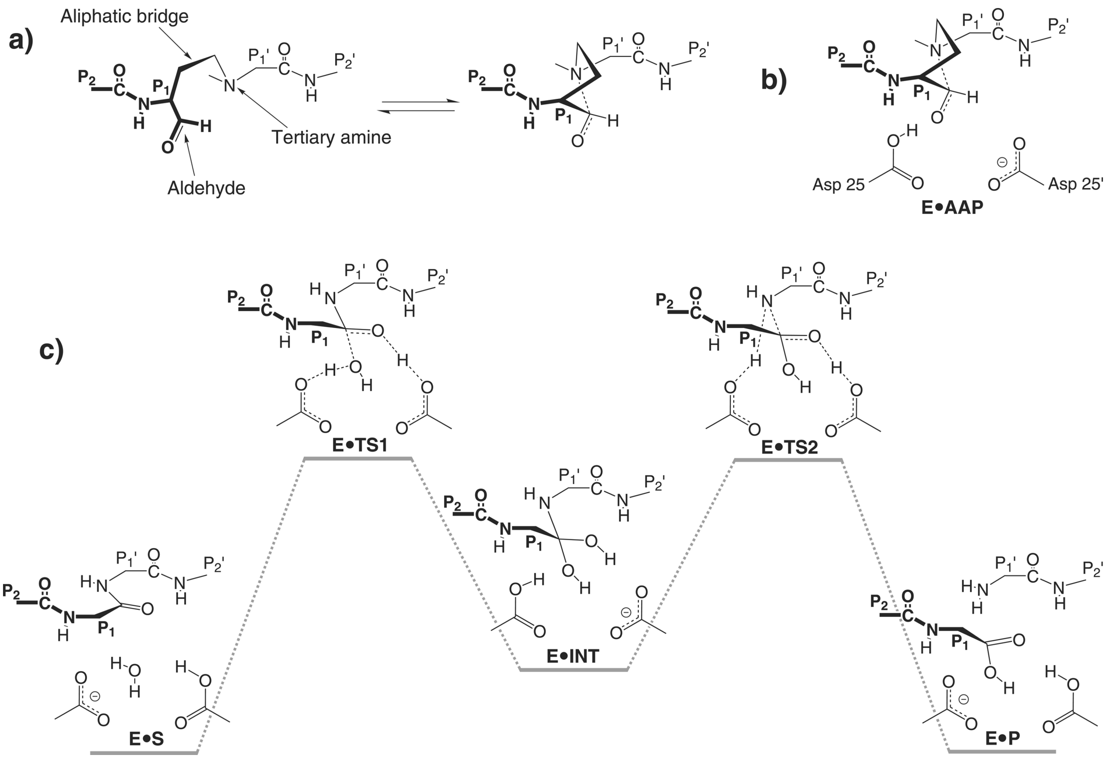


*Figure 2.* N···CO bond as a transition state mimic. (a) N···CO bond formation within an amino-aldehyde peptide (AAP). Side chains of the AAP are not represented for the sake of simplicity. Instead, their positions are indicated using the notation of Schechter and Berger ($P_2$, $P_1$, $P_1'$, $P_2'$).[84] The N···CO core is introduced at the $P_1-P_1'$ junction, the proximity of the amine and aldehyde fragments being ensured by an *aliphatic bridge*. (b) HIV-1 PR aspartyl dyad complexed with an AAP. (c) Catalytic mechanism of HIV-1 PR involving a tetrahedral intermediate (E•INT), which is represented here as a gem-diol. Note that some authors have reported that this intermediate could be an oxyanion (see text).[41,42]

compounds in the gas phase—provided evidence for a charged oxyanion.[41] Other authors calibrated their EVB Hamiltonian without including the gem-diol in their resonance structure set.[42] Since we have chosen a computational setup involving an *ab initio* approach similar to that of ref 39, we have considered a neutral gem-diol in the present study. However, we stress that this choice should not affect our results dramatically, since the tetrahedral intermediate is just chosen as a starting structure to perform a first equilibration of the system. More specifically, we chose the complex between HIV-1 PR (E) and the Thr−Ile−Met−Met−Gln−Arg peptide substrate[43] in its hydrated form (INT; Figure 2c). E•INT was constructed from the X-ray structure of HIV-1 PR complexed with the highly peptidic MVT-101 inhibitor[34,36] (4HVP entry in the PDB data bank[44]), a compound that exhibits the sequence N-acetyl-Thr-Ile-Nle-Ψ[$CH_2NH$]-Nle-Gln-Arg-amide (Nle = norleucine). The Asp dyad was assumed to be monoprotonated according to the commonly accepted mechanism of HIV-1 PR.[36,39,41,42,45] The protein was immersed in a 90 × 71 × 74 $Å^3$ water box, and the entire system was neutralized by adding six chloride counterions. The whole system was composed of about 38 000 atoms.

**2.2. Equilibration of E•INT.** *2.2.1. Classical MD Simulations.* The complex was first equilibrated at the classical level using the AMBER9 suite of programs.[46] The parameters for the solute, apart from the amide hydrate (−$C(OH)_2$−NH−) moiety of the gem-diol intermediate, were taken from the AMBER 03 force field,[47] and the TIP3P model was used to describe water molecules.[48] Bonded and van der Waals parameters of the amide hydrate were extracted from the generalized AMBER force field (GAFF).[49] The charges of the amide hydrate were obtained using a standard RESP procedure.[50] The electrostatic potential was computed at the HF/6-31G(d) level with the Gaussian 03 package,[51] from a model compound including the hydrated Met−Met sequence capped with acetyl and N-methyl groups, i.e. Ace−Met−[$C(OH)_2$−NH]−Met−Nme. Note that these additional parameters are used for a small part only of the entire system, which is then described within the QM part during further equilibration at the QM/MM level (see next section).

Long-range electrostatic interactions were computed using the Ewald particle mesh method.[52,53] A cutoff of 10 Å was applied for the van der Waals interactions and the real part of the electrostatic interactions. A time step of 1.5 fs was used, and all bonds containing hydrogen were constrained using the SHAKE algorithm. Constant temperature was achieved using Langevin dynamics[54] with a collision frequency of 5 $ps^{-1}$, while the pressure was maintained using a Berendsen's barostat[55] with a relaxation time of 1.0 ps. The system was first heated to 150 K over 15 ps and then to 300 K over a further 15 ps. Then, an equilibration of 1 ns at 1 atm and 300 K was carried out.

*2.2.2. QM/MM MD Simulations.* Starting from the equilibrated E•INT structure obtained at the classical level, we switched to a hybrid quantum mechanics/molecular mechanics (QM/MM)[28] description for the system. We used the

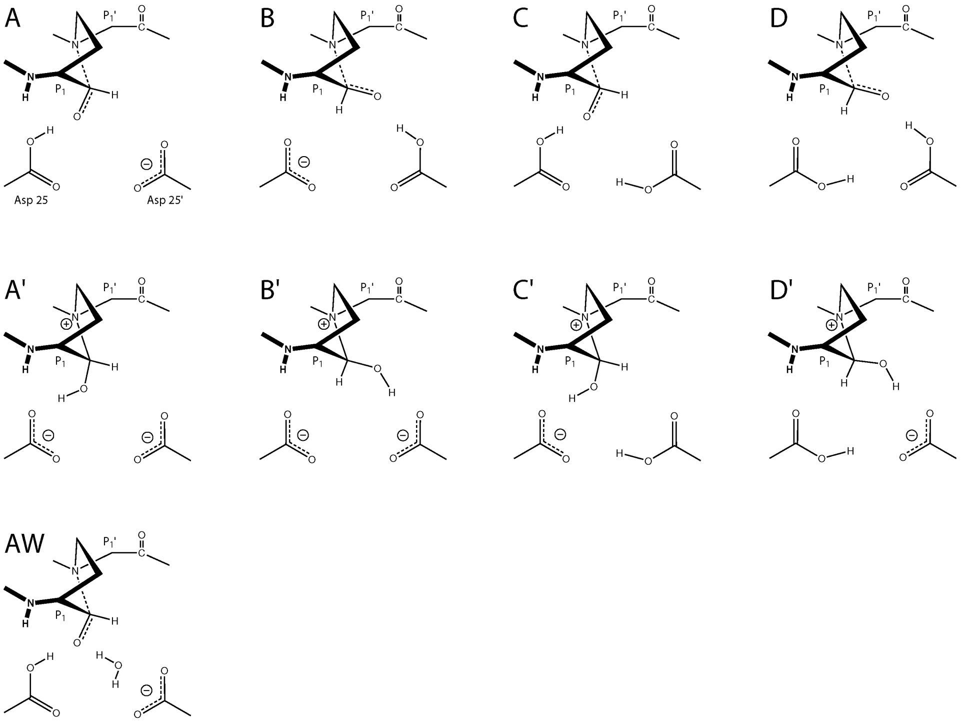


***Figure 3.*** E•AAP isomers and protonation states considered in this study.

approach developed by Rothlisberger and co-workers.[56−58] The QM region encompassed the Asp25(25′) side chains and the amide hydrate moiety and was described at the DFT/BLYP[59,60] level of theory. Dangling bonds were saturated with hydrogen atoms. The Kohn−Sham orbitals were expanded in plane waves with a cutoff of 70 Ry and a quantum cell of 17.2 × 14.8 × 14.8 $Å^3$. A fictitious electron mass of 600 au and a time step of 5 au (≈ 0.12 fs) were used.

The E•INT system was first minimized using a simulated annealing-like procedure: starting from a temperature of 50 K, atomic velocities were rescaled at each time step by a factor of 0.99. Then, the system was heated up to 300 K over 15 ps. Subsequently, 12 ps of NVT simulation were performed, using a Nosé−Hoover chain thermostat[61−63] of 900 $cm^{-1}$ frequency.

**2.3. Transformation of E•INT into E•AAP.** A structure obtained after 10.3 ps of simulation was chosen to perform the E•INT → E•AAP transformation. The hydrated substrate was modified at the Met−Met junction to incorporate the N···CO core and the aliphatic bridge (Figure 2b). The starting value of the characteristic C−N distance of the N···CO moiety was chosen to be 1.7 Å, which is a typical value for a N···CO bond in a polar-protic medium.[33] At this stage, we should stress that the only groups that differ from the initial crystallographic structure are the N···CO core and the aliphatic bridge. Once the peptide backbone of the AAP is accommodated in the active site, there are few possibilities left for the positioning of the cycle made of the aliphatic bridge and the N···CO group. The main uncertainty lies in the configuration of the carbon atom of the N···CO moiety. Thus we decided to study both possible configurations (Figure 3A,B).

Once the starting position of heavy atoms was chosen, we had to address the question of the active site protonation state within the E•AAP complex. Indeed, HIV-1 PR exhibits a wide range of protonation patterns according to the presence and the nature of the ligand.[39,45,64−71] The prediction of such a pattern can be done, for instance, by fitting kinetic data recorded at different pH to rate equations,[45,64] by computational $pK_a$ estimation from an experimental structure,[65,66] by NMR titration,[67,68] by constructing computational models that aim at reproducing experimental data such as crystallographic structures[70,71] or NMR chemical shifts,[69] or by combining X-ray and neutron crystallography.[72]

The *a priori* determination of the E•AAP active-site protonation pattern is challenging. We decided to apply a systematic strategy, in which a series of protonation states for each configuration of the N···CO carbon atom was considered. When bound to neutral ligands, the aspartyl dyad of HIV-1 PR is usually monoprotonated[39,45,71,73] or diprotonated,[67] while positively charged ligands may yield an unprotonated dyad.[45,65]

Following this, we generated a series of nine isomers in different protonation states that are depicted in Figure 3.

First, we considered complexes A,B (monoprotonated dyad with neutral AAP) and C,D (diprotonated dyad with neutral AAP). In addition, the N···CO interaction can also be described by the limiting $N^+−C−O^−$ form. The negatively charged oxygen atom can be seen as an alcoholate and can thus be very basic. We have thus considered the possible proton transfer from the neighboring carboxyl group to the N···CO core leading to a positively charged AAP containing a $N^+−CO−H$ moiety. A′,B′ (unprotonated dyad with positively charged AAP) and C′,D′ (monoprotonated dyad with positively charged AAP) were obtained from A,B and C,D, respectively, by shifting the closest proton of the Asp dyad to the N···CO oxygen atom.

To our knowledge, the catalytic water molecule that is tightly H-bonded to the aspartyl dyad in the HIV-1 PR-substrate complex (Figure 2c) is systematically displaced by any active site inhibitor bound to the enzyme. Indeed, it has only been observed in crystallographic structures of the free enzyme,[74,75] but never in any X-ray or NMR structure of the bound enzyme. However, we found it interesting to check if an AAP can displace this water molecule. Thus, we considered a last complex, denoted by AW, in which a water molecule is added close to the aspartyl dyad.

**2.4. QM/MM Modeling of E•AAP Isomers in Different Protonation States.** Once the E•INT → E•AAP transformation was performed, each E•AAP structure underwent a mild annealing-like protocol allowing a slow relaxation and minimization of the newly introduced aliphatic bridge−N···CO moiety and the surrounding protein medium. This enabled us to scrutinize (i) the stability of the N···CO bond within the protein and (ii) the interactions between the AAP and the protein.

We used the same level of calculation as the one we used to equilibrate E•INT. The QM region included the Asp25(25′) side chains, the N···CO moiety, and the aliphatic bridge. We stress that the ability of density functional theory to describe the N···CO interaction was demonstrated previously.[33] The quantum cell size was adapted to that of the QM region, leading to a 16.9 × 14.3 × 16.9 Å$^3$ box.

Each E•AAP complex underwent a first minimization, using the same simulated annealing-like procedure that we used previously to optimize the geometry of the E•INT complex, i.e., using a starting temperature of 50 K and a scaling factor of 0.99. These calculations were stopped as soon as the temperature reached a value below 3 K. Then, the system was let free to relax during 8 ps of NVE QM/MM molecular dynamics. During this run, each complex heated up to about 40 K due to remaining bad contacts. Finally, the minimization of the system was achieved using a second annealing with a scaling factor of 0.999.

To check the validity of this relaxation/optimization protocol, we performed further calculations using constraints. These calculations are described in the Supporting Information.

**2.5. QM/MM Modeling of AAP in Water.** The simulations in water were started from the same starting structure of the AAP as the one used in complex A. The inhibitor was immersed in a 49 × 49 × 49 Å$^3$ water box equilibrated at room temperature at the classical level. All of the atoms of the AAP were fixed during the first stages of the simulation. For the QM/MM runs, we used the same description as the one for the E•AAPs. The quantum cell size was adapted to that of the QM region, leading to a 11.0 × 11.3 × 10.4 Å$^3$ box.

The water box was further equilibrated using the following protocol: A first annealing was performed using a starting temperature of 50 K and a velocity scaling factor of 0.99. Then, the solvent underwent 4.7 ps of MD simulation at 300 K using the Berendsen weak coupling algorithm.[55] Finally, the target temperature was decreased linearly from 300 to 1 K in 3 ps.

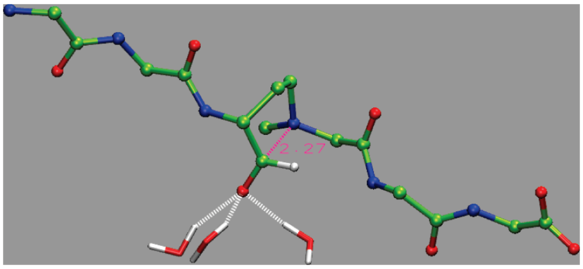


*Figure 4.* Optimized geometries of the AAP in water at the QM/MM level of theory. Only the inhibitor backbone and water molecules hydrogen bonded to the oxygen atom of the N···CO core are represented. The N···CO interaction is depicted in magenta.

At this stage, the constraints on the inhibitor were removed, and we minimized the QM and the MM part of the system consecutively. Then, the system underwent 4 ps of NVT simulation using a Nosé−Hoover chain thermostat.[61−63] Finally, a minimized geometry was obtained by performing an annealing with a velocity scaling factor of 0.999.

## 3. Results and Discussion

**3.1. AAP in Water.** We start the analysis of our results by describing the behavior of the N···CO bond embedded in an AAP in an aqueous medium. During our 4 ps of NVT simulation at 300 K, a weak N···CO interaction was maintained, with an average value of the C−N distance of 2.41 Å and a standard deviation of 0.18 Å. After minimization using our annealing-like protocol, this distance decreased to 2.27 Å. Figure 4 shows the corresponding optimized geometry. The nitrogen lone pair of the N···CO core is directed toward the aldehyde, which has lost its coplanarity. Three water molecules are hydrogen bonded to the aldehyde, which stabilizes the N···CO bond.

The fact that the N···CO bond is maintained over the course of the simulation shows that our computational approach is able to reproduce the closed configuration reported by Hasserodt et al.[19] On the basis of NMR measurements in methanol, they estimated that 70% of the AAPs exhibit a N···CO bond. We expect that much longer simulations would provide opening and closing events.

Nevertheless, the average C−N distance we observe is longer than the typical value of 1.8 Å of a N···CO bond in a polar-protic medium.[33] This is in agreement with the computational study of Pilmé et al., who have shown that water molecules H-bonded to the N···CO core stabilize the N$^+$−C−O$^-$ form by accepting part of its electronic density.[33] Even though the charge transfer is small (ca. 0.06 e/molecule), it was shown to be sufficient to stabilize short CN distances. In our simulations, the interaction between the N···CO core and the surrounding water molecules is described through the QM/MM interface, which does not account for charge transfer effects, hence leading to a longer C−N distance. Note, however, that this problem does not hold for our simulations in HIV-1 PR (next sections), because the polar protein group close to the N···CO core was included in the QM part.

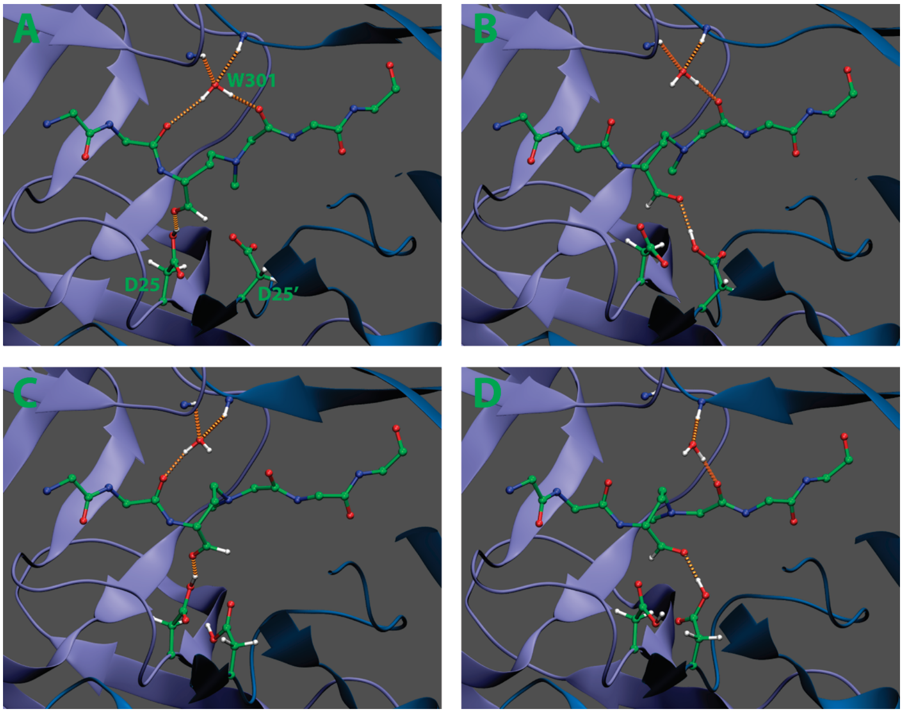


***Figure 5.*** Optimized geometries of complexes A−D at the QM/MM level of theory. Only polar hydrogens belonging to the H-bond networks around W301, the N···CO core, and the aspartyl dyad are represented.

***Table 1.*** Main Geometrical Parameters Resulting from the Minimization of Complexes A−D[a]

| | A | B | C | D |
|---|---|---|---|---|
| $d_{\mathrm{CN}}$[b] | 3.75 | 3.43 | 3.01 | 2.51 |
| $\mathrm{NCO\cdots H\delta_{Asp25(25')}}$[c] | 1.66 | 1.76 | 1.60 | 1.61 |
| $\mathrm{NH_{Ile50}\cdots O_{W301}}$[d] | 2.20 | 2.39 | 2.34 | 3.42 |
| $\mathrm{NH_{Ile50'}\cdots O_{W301}}$[d] | 2.26 | 2.01 | 2.08 | 1.87 |
| $\mathrm{CO_{P2}\cdots H_{W301}}$[d] | 2.30 | 3.67 | 1.77 | 4.28 |
| $\mathrm{CO_{P1'}\cdots H_{W301}}$[d] | 1.75 | 1.67 | 3.83 | 1.64 |
| $\Omega$[e] | 94.25 | 69.44 | 37.77 | −43.12 |
| $\Phi$[f] | −125.77 | −84.81 | −74.34 | −51.06 |

[a] Distances are given in Å and angles in degrees. [b] Distance between the tertiary amine nitrogen and the aldehyde. [c] Hydrogen bond between the N···CO oxygen and the $H_\delta$ of the aspartyl dyad. Residue 25 or 25′ is considered, depending on the protonation pattern of the aspartyl dyad. [d] Hydrogen bond network around the structural water molecule W301. [e] Dihedral angle between each Asp oxygen of the aspartyl dyad. [f] Dihedral angle of the aliphatic bridge.

**3.2. Protonation State of the N···CO Oxygen Atom in the Protein.** The first question we addressed concerning the behavior of AAPs within HIV-1 PR was whether the protein is able to provide a stabilizing medium for a protonated, positively charged AAP. Indeed, due to the enhanced ionic character of the C−O bond within the N···CO core, the basicity of the oxygen is increased, and thus, a sufficiently acidic medium could stabilize a covalent $\mathrm{N^+{-}CO{-}H}$ moiety.

In the early stages of the minimization of isomers A′−D′, we observed that the proton linked to the N···CO oxygen was systematically shifted to the closest aspartate, leading to isomers A−D, respectively. Thus, despite the difficulty of providing an a priori accurate estimation of the protonation pattern in the active site of our system, our calculations consistently converge to an AAP N···CO core that is unprotonated within HIV-1 PR. Therefore, in the following, only isomers A−D and AW will be considered.

**3.3. N···CO Bond Stability in the Active Site and Interaction with the Asp Dyad.** Optimized geometries of complexes A−D are depicted in Figure 5. Selected structural parameters are reported in Table 1. For each complex, the characteristic C−N distance of the N···CO moiety exhibits a drastic lengthening from the starting value of 1.7 Å. Note that, since this opening event was already observed during geometry optimization, we did not attempt to run molecular dynamics simulations at room temperature, as we did for the modeling of AAPs in water. Instead, we applied a mild relaxation/minimization protocol in order to assess the effect of the enzymatic media on the geometry of AAPs.

Monoprotonated complexes A and B are stabilized in a fully opened conformation, with $d_{\mathrm{CN}} = 3.75$ and 3.43 Å in A and B, respectively. This distance is larger than the characteristic value of even a weak N···CO interaction in which $d_{\mathrm{CN}} \approx 2.8$ Å.[33] In both structures, the N···CO nitrogen lone pair is no longer directed toward the planar aldehyde group, indicating that the N···CO interaction is completely disrupted. Furthermore, the aspartyl dyad coplanarity is lost according to the values of the dihedral angle between each Asp oxygen, i.e., $\Omega = 94.2$ and 69.4° for A and B, respectively. This structural feature plays a crucial role in the binding of both substrate and inhibitors.[4,76,77] The highly distorted conformation of the aspartyl dyad is a clear indicator of an unfavorable interaction between the N···CO core, Asp25 and Asp25′. Despite the hydrogen bond between the proton of Asp25(25′) and the carbonyl oxygen of the AAP in isomer A(B) (see Table 1), the AAP does not provide sufficient shielding to stabilize the strong electrostatic Asp−Asp repulsion.[70]

The AW complex (results not shown) follows a similar route, leading to a drastic lengthening of the C−N distance and the loss of the aspartyl dyad coplanarity. Moreover, the opening of the N···CO core is accompanied by the departure of the catalytic water molecule out of the aspartyl dyad, yielding a structure close to complex A. In fact, we observe that this water molecule is stabilized by a hydrogen bond with the Gly27 carbonyl group. Despite the short time scale that is accessible at the QM/MM level, we expect that this is a transient state, prior to a move of the catalytic water molecule to the bulk. We have already observed this behavior in a classical simulation of a HIV-1 PR-substrate complex (not presented here), which is consistent with observations reported by others.[40]

Diprotonated complex D exhibits the smallest C−N distance ($d_{CN}$ = 2.51 Å), which corresponds to a weak N···CO interaction.[33] The nitrogen lone pair is still directed toward the carbonyl group which is less planar and more tightly H-bound to Asp25′ than in the monoprotonated complexes, according to the NCO···$H\delta_{Asp25'}$ distance. The Asp dyad remains almost coplanar, suggesting that the complex is much more stable than the monoprotonated ones. This is in agreement with the fact that the shielding introduced by the additional proton in diprotonated states makes the aspartyl dyad more amenable to accepting the accumulated negative charge of the N···CO oxygen.

Complex C represents an intermediate situation between complexes D and A,B. As in D, it is characterized by a coplanar aspartyl dyad and a tight hydrogen bond between the oxygen of the N···CO core and Asp25. However, similarly to A and B, the C−N distance equals 3.01 Å, which corresponds at best to a shallow N···CO bond.

Our systematic approach reflects the geometric behavior of the N···CO core within the enzyme, over a set of different conditions related to the protonation state and the starting configuration of the N···CO carbon atom. As we will show in the next section, our model enables one to establish a correlation between the disruption of the N···CO bond and other interactions that play a key role in the affinity between the AAP and HIV-1 PR.

**3.4. Origin of the N···CO Opening.** A detailed analysis of the hydrogen bond network inside the HIV-1 PR active site sheds some light on the origin of the instability of the N···CO bond within the enzyme. In HIV-1 PR−substrate and in most HIV-1 PR−inhibitor complexes,[78] a tetrahedrally coordinated structural water molecule, commonly labeled W301,[1,79,80] bridges Ile50(50′) NH groups belonging to the upper part of the active site cleft (the so-called *flaps*, Figure 1) and $P_2$ and $P_1'$ CO groups of the substrate/inhibitor. Hence, this hydrogen bond network plays a crucial role in the correct positioning of both substrate and peptidomimetic inhibitors in the active site. Inhibitors that do not exhibit these interactions are those that have been designed to displace W301, such as cyclic urea derivatives[26] or the FDA-approved drug Tipranavir.[27]

Figure 6a depicts the starting geometry (just after the E•INT→E•AAP transformation) of W301 and its surrounding atoms within complex A. Note that the position of heavy atoms is the same as that of complex C and AW and is very similar to that of complexes B and D, the only difference being the configuration of the N···CO carbon atom. The tetrahedral H-bond network around W301 is present, prior to any step of our minimization protocol. In the starting configuration of our complexes, the “macrocycle” composed of W301, $P_2$, and $P_1'$ CO groups and the aliphatic bridge (represented with orange transparent tubes in Figure 6) is in a quite compact conformation, while the C−N distance of the N···CO moiety is 1.7 Å.

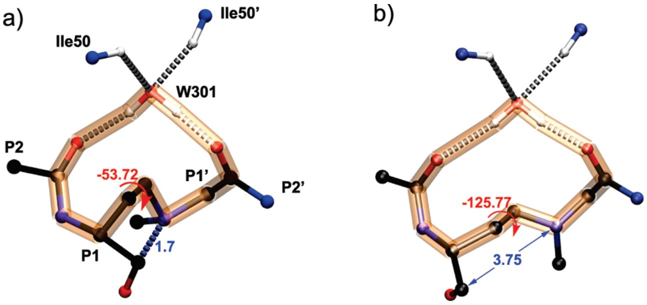


***Figure 6.*** Competition between the N···CO bond formation and the conservation of the H-bond network around the structural water molecule W301. Starting (a) and final (b) geometries of W301 and its surrounding atoms within complex A. Side chains and hydrogen atoms not belonging to the H-bond network around W301 are not represented for the sake of simplicity. The “macrocycle” composed of W301, $P_2$, and $P_1'$ CO groups and the aliphatic bridge is represented with orange transparent tubes. The H bonds around W301 are depicted with gray-dashed tubes, while the N···CO interaction is represented with a blue-dashed tube in the starting structure (a).

During the NVE molecular dynamics run of complex A, the H bonds around W301 are first partly disrupted, and as the C−N distance lengthens, the hydrogen bond network is progressively restored. Figure 6b represents the final (optimized) structure of complex A. The “macrocycle” is stabilized in an extended conformation, in which the aliphatic bridge has been pushed away from W301. This structural reorganization occurs together with a drastic increase of both the characteristic dihedral angle Φ of the aliphatic bridge (from −51 to −126°) and the C−N distance (from 1.7 to 3.75 Å). Hence, the disruption mechanism of the N···CO bond observed in complex A may be formulated as follows: The hydrogen bond network that W301 tends to form with two backbone carbonyl groups of the AAP tightens the “macrocycle”, which in turn reduces the steric hindrance by extending the aliphatic bridge, thus lengthening the C−N distance of the N···CO core.

Complex B exhibits a behavior similar to that of complex A, sharing the same location of W301 and the evolution of the structure toward an extended “macrocycle”. The major difference lies in the value of the dihedral angle of the aliphatic bridge (Table 1), which is lower in B, i.e., Φ = −84.8°. Thus, the aliphatic bridge remains closer to the center of the “macrocycle”, and W301 cannot establish an optimal hydrogen bond with the $P_2$ carbonyl.

In order to analyze the link between the hydrogen bond network around W301 and the CN distance, we have conducted a constrained optimization. Starting from the optimized geometry of complex B, we changed the hydrogen bond network around the W301 molecule from that in complex B to that in complex A. The CN bond was not frozen, and we found that it increased to reach a final length of 3.70 Å, close to the value observed in complex A.

The situation is rather different for diprotonated complexes. In the optimized complexes C and D, W301 is no longer tetrahedrally coordinated, as depicted in Figure 5C,D. The highest disruption occurs for complex D, which conserves only two hydrogen bonds. On the other hand, the aliphatic bridge remains in a conformation closer to that of the starting structure, as indicated by the $\Phi$ values: $-74.3$ and $-51.1°$ for C and D, respectively. This is consistent with the C−N distance values discussed in the previous section. As previously noted, a diprotonated dyad is a better hydrogen bond donor than a monoprotonated one and is thus more favorable to the N•••CO interaction, which in turn, is more competitive against the conservation of the H-bond network around W301.

The behavior observed for complexes A−D can be summarized stating that a N•••CO bond and a proper H-bond network around W301 cannot be realized both at the same time. Clearly, the introduction of the N•••CO core at the scissile peptide bond location induces a systematic competition between the N•••CO bond formation and the interaction network involving the structural water molecule W301. Barillari et al.[81] estimated the binding free energy of W301 to a series of complexes between HIV-1 PR and peptidomimetic inhibitors using the double-decoupling free energy simulation method. They found that the binding free energy ranged from $-7$ to $-10$ kcal mol$^{-1}$. If one removes the entropy contribution, in which the upper bound has been estimated to be about 2 kcal mol$^{-1}$,[82] one obtains a binding energy of $-(9-12)$ kcal mol$^{-1}$. This is comparable to the energy of the N•••CO bond, which has been estimated to be $-11$ kcal mol$^{-1}$ at the CCSD(T) level.[33] These energetic considerations support the competition that we observed between the conservation of the hydrogen bond network around W301 and N•••CO bond formation.

Since the N•••CO moiety has been designed to interact strongly with the aspartyl dyad,[19] our simulations show that an AAP cannot provide stabilizing interactions with all the key parts of the HIV-1 PR binding site at the same time. The low inhibition power of AAPs might originate, at least partly, from this competition. Interestingly, our results suggest that further development of N•••CO containing inhibitors should focus on nonpeptidic compounds that could displace W301. This in line with the hydrazino-urea compounds proposed recently by Hasserodt et al.[20,21] These derivatives contain both a N•••CO bond and a hydrazino-urea group designed to interact directly with the flaps of HIV-1 PR, similar to cyclic urea derivatives[26] and the FDA-approved drug Tipranavir.[27]

Note that hydrazino−urea compounds synthesized by Hasserodt et al. bind only slightly stronger to HIV-1 PR than AAPs ($K_i \approx 29$ $\mu$M and 97 $\mu$M, respectively). However, the former contain only three groups aimed at filling the subpockets of the enzyme binding cleft, while cyclic urea derivatives usually have four.[26] Hence, we encourage the development of optimized N•••CO-containing hydrazino−urea inhibitors that would contain proper groups aimed at interacting with the same binding subpockets as cyclic urea derivatives. Such a design could be assisted by a computational study based on the same approach we developed in the present work.

## 4. Conclusions

We have proposed here a computational procedure to tackle the difficult theoretical description of HIV-1 PR inhibitors based on the unusual N•••CO bond. This procedure consists of an explicitly solvated model of the ligand−enzyme complex (E•AAP) described at the atomistic level with a QM/MM approach. The N•••CO core and part of the enzyme active site are described using an accurate QM level, while the rest of the protein and the solvent are described using the classical AMBER force field.

In this work, we have applied this model to investigate the origin of the low inhibiting power of the recently proposed amino-aldehyde peptide (AAP) compounds against HIV-1 PR. Our calculations provide detailed information on the feasibility of the N•••CO bond formation within the enzymatic environment along with crucial interactions that govern the stability of the complex.

Considering all the possible protonation patterns of the active site aspartyl dyad, we have shown that N•••CO bond formation/dissociation takes place in a competitive mechanism, in which the structural water molecule W301 tends to establish a hydrogen bond network that indirectly penalizes the shortening of the distance between the nitrogen atom and the CO group of the N•••CO core. We conclude that the reported poor inhibition power of AAPs[19] originates, at least partly, from this competition.

Despite this, a N•••CO•••H$\delta_{\mathrm{Asp25(25')}}$ hydrogen bond was observed for each protonation state. In the case of complex D, this H bond is tighter, and a weak N•••CO interaction is formed at the same time. This means that, under appropriate conditions, a N•••CO core could interact strongly with the aspartyl dyad of HIV-1 PR. Hence, the design of N•••CO-containing candidates that could displace the water molecule W301 would be an interesting alternative to AAPs. This supports the idea that a hydrazino−urea core[20,21] is an interesting template for the design of potent anti-AIDS drugs. In particular, it would be interesting to extend the recent work of Hasserodt et al. to hydrazino−urea derivatives containing peripheral groups aimed at filling the HIV-1 PR subpockets $P_2$, $P_1$, $P'_1$, $P'_2$ properly, similar to cyclic urea derivatives. Such a design could be done *in silico*, using the approach we have developed in the present study.

**Acknowledgment.** We thank the "Centre Informatique National de l'Enseignement Supérieur" and the "Institut du développement et des ressources en informatique Scientifique" for computational support and resources. We also thank Dr. M. Waibel and Pr. J. Hasserodt for useful discussions. The financial support from the "Cluster de recherche Chimie de la Région Rhone-Alpes" is duly acknowledged.

## Abbreviations

HIV-1, Human Immunodeficiency Virus Type 1; AIDS, Acquired ImmunoDeficiency Syndrome; PR, Protease; E, catalytically active dimeric form of HIV-1 protease; INT, Thr−Ile−Met−Met−Gln−Arg peptide substrate in its hydrated form; QM/MM, Quantum Mechanics/Molecular Mechanics; FDA, Food and Drug Administration; AAP, Amino-Aldehyde Peptide; CPMD, Car−Parrinello Molecular Dynamics; DFT, Density Functional Theory.

**Supporting Information Available:** The results of the additional simulations with various constrained distances ($d_{CN}$, $NH_{Ile50}\cdots O_{W301}$, $NH_{Ile50'}\cdots O_{W301}$, $CO_{P2}\cdots H_{W301}$, $CO_{P1'}\cdots H_{W301}$) are described. This material is available free of charge via the Internet at http://pubs.acs.org.

## References

(1) Wlodawer, A.; Vondrasek, J. *Annu. Rev. Biophys. Biomol. Struct.* **1998**, *27*, 249–284.

(2) Wlodawer, A.; Erickson, J. W. *Annu. Rev. Biochem.* **1993**, *62*, 543–585.

(3) Debouck, C.; Gorniak, J. G.; Strickler, J. E.; Meek, T. D.; Metcalf, B. W.; Rosenberg, M. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 8903–8906.

(4) Fitzgerald, P. M. D.; Springer, J. P. *Annu. Rev. Biophys. Biophys. Chem.* **1991**, *20*, 299–320.

(5) Poorman, R.; Tomasselli, A.; Heinrikson, R.; Kezdy, F. *J. Biol. Chem.* **1991**, *266*, 14554–14561.

(6) Kohl, N. E.; Emini, E. A.; Schleif, W. A.; Davis, L. J.; Heimbach, J. C.; Dixon, R. A. F.; Scolnick, E. M.; Sigal, I. S. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 4686–4690.

(7) Seelmeier, S.; Schmidt, H.; Turk, V.; von der Helm, K. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 6612–6616.

(8) McQuade, T. J.; Tomasselli, A. G.; Liu, L.; Karacostas, V.; Moss, B.; Sawyer, T. K.; Heinrikson, R. L.; Tarpley, W. G. *Science* **1990**, *247*, 454–456.

(9) Richman, D. D. *Nature* **2001**, *410*, 995–1001.

(10) Hammer, S. M.; Squires, K. E.; Hughes, M. D.; Grimes, J. M.; Demeter, L. M.; Currier, J. S.; Eron, J. J.; Feinberg, J. E.; Balfour, H. H.; Deyton, L. R.; Chodakewitz, J. A.; Fischl, M. A.; Phair, J. P.; Pedneault, L.; Nguyen, B.-Y.; Cook, J. C. *N. Engl. J. Med.* **1997**, *337*, 725–733.

(11) Gulick, R. M.; Mellors, J. W.; Havlir, D.; Eron, J. J.; Gonzalez, C.; McMahon, D.; Richman, D. D.; Valentine, F. T.; Jonas, L.; Meibohm, A.; Emini, E. A.; Chodakewitz, J. A.; Deutsch, P.; Holder, D.; Schleif, W. A.; Condra, J. H. *N. Engl. J. Med.* **1997**, *337*, 734–739.

(12) Condra, J. H.; Schleif, W. A.; Blahy, O. M.; Gabryelski, L. J.; Graham, D. J.; Quintero, J.; Rhodes, A.; Robbins, H. L.; Roth, E.; Shivaprakash, M.; Titus, D.; Yang, T.; Tepplert, H.; Squires, K. E.; Deutsch, P. J.; Emini, E. A. *Nature* **1995**, *374*, 569–571.

(13) Martinez-Cajas, J. L.; Wainberg, M. A. *Antiviral Res.* **2007**, *76*, 203–221.

(14) Luque, I.; Todd, M. J.; Gómez, J.; Semo, N.; Freire, E. *Biochemistry* **1998**, *37*, 5791–5797.

(15) Velázquez-Campoy, A.; Kiso, Y.; Freire, E. *Arch. Biochem. Biophys.* **2001**, *390*, 169–175.

(16) Velázquez-Campoy, A.; Luque, I.; Freire, E. *Thermochim. Acta* **2001**, *380*, 217–227.

(17) Vega, S.; Kang, L.-W.; Velázquez-Campoy, A.; Kiso, Y.; Amzel, M.; Freire, E. *Proteins: Struct., Funct., Bioinf.* **2004**, *55*, 594–602.

(18) Leung, D.; Abbenante, G.; Fairlie, D. P. *J. Med. Chem.* **2000**, *43*, 305–341.

(19) Gautier, A.; Pitrat, D.; Hasserodt, J. *Bioorg. Med. Chem.* **2006**, *14*, 3835–3847. Kinetics and inhibition measurements are given in their Supplementary Data.

(20) Waibel, M.; Hasserodt, J. *J. Org. Chem.* **2008**, *73*, 6119–6126.

(21) Waibel, M.; Pitrat, D.; Hasserodt, J. *Bioorg. Med. Chem.* **2009**, *17*, 3671–3679.

(22) Shuman, C. F.; Hämäläinen, M. D.; Danielson, U. H. *J. Mol. Recognit.* **2004**, *17*, 106–119.

(23) Velázquez-Campoy, H. O. A.; Xie, D.; Freire, E. *Protein Sci.* **2002**, *11*, 1908–1916.

(24) Markgren, P.-O.; Schaal, W.; Hämäläinen, M.; Karlén, A.; Hallberg, A.; Samuelsson, B.; Danielson, U. H. *J. Med. Chem.* **2002**, *45*, 5430–5439.

(25) Velázquez-Campoy, A.; Freire, E. *J. Cell. Biochem.* **2001**, *37*, 82–88.

(26) Lam, P. Y. S.; Ru, Y.; Jadhav, P. K.; Aldrich, P. E.; DeLucca, G. V.; Eyermann, C. J.; Chang, C.-H.; Emmett, G.; Holler, E. R.; Daneker, W. F.; Li, L.; Confalone, P. N.; McHugh, R. J.; Han, Q.; Li, R.; Markwalder, J. A.; Seitz, S. P.; Sharpe, T. R.; Bacheler, L. T.; Rayner, M. M.; Klabe, R. M.; Shum, L.; Winslow, D. L.; Kornhauser, D. M.; Jackson, D. A.; Erickson-Viitanen, S.; Hodge, C. N. *J. Med. Chem.* **1996**, *39*, 3514–3525.

(27) Doyon, L.; Tremblay, S.; Bourgon, L.; Wardrop, E.; Cordingley, M. G. *Antiviral Res.* **2005**, *68*, 27–35.

(28) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227–249.

(29) Spanka, G.; Boese, R.; Rademacher, P. *J. Org. Chem.* **1987**, *52*, 3362–3367.

(30) Griffith, R.; Bremner, J. B.; Titmuss, S. J. *J. Comput. Chem.* **1997**, *18*, 1211–1221.

(31) Leonard, N. J.; Oki, M. *J. Am. Chem. Soc.* **1954**, *76*, 3463–3465.

(32) Kirby, A. J.; Komarov, I. V.; Bilenko, V. A.; Davies, J. E.; Rawson, J. M. *Chem. Commun.* **2002**, 2106.

(33) Pilmé, J.; Berthoumieux, H.; Robert, V.; Fleurat-Lessard, P. *Chem.*−*Eur. J.* **2007**, *13*, 5388–5393.

(34) Miller, M.; Schneiderand, J.; Sathyanarayana, B.; Tothand, M.; Marshalland, G.; Clawsonand, L.; Selkand, L.; Kent, S.; Wlodawer, A. *Science* **1989**, *246*, 1149–1152.

(35) Prabu-Jeyabalan, M.; Nalivaika, E.; Schiffer, C. A. *Structure* **2002**, *10*, 369–381.

(36) Hyland, L. J.; Tomaszek, T. A., Jr.; Roberts, G. D.; Carr, S. A.; Magaard, V. W.; Bryan, H. L.; Michael, S. A. F.; Moore, L.; Minnich, M. D.; Culp, J. S.; DesJarlais, R. L.; Meek, T. D. *Biochemistry* **1991**, *30*, 8441–8453.

(37) Kovalevsky, A. Y.; Chumanevich, A. A.; Liu, F.; Louis, J. M.; Weber, I. T. *Biochemistry* **2007**, *46*, 14854–14864.

(38) Kumar, M.; Prashar, V.; Mahale, S.; Hosur, M. V. *Biochem. J.* **2005**, *389*, 365–371.

(39) Piana, S.; Bucher, D.; Carloni, P.; Rothlisberger, U. *J. Phys. Chem. B* **2004**, *108*, 11139–11149.

(40) Carnevale, V.; Raugei, S.; Piana, S.; Carloni, P. *Comput. Phys. Commun.* **2008**, *179*, 120–123.

(41) Trylska, J.; Grochowski, P.; McCammon, J. A. *Protein Sci.* **2004**, *13*, 513–528.

(42) Bjelic, S.; Aqvist, J. *Biochemistry* **2006**, *45*, 7709–7723.

(43) Ratner, L.; Haseltine, W.; Patarca, R.; Livak, K. J.; Starcich, B.; Josephs, S. F.; Doran, E. R.; Rafalski, J. A.; Whitehorn, E. A.; Baumeister, K.; Ivanoff, L.; Petteway, S. R., Jr.; Pearson, M. L.; Lautenberger, J. A.; Papas, T. S.; Ghrayeb-parallel, J.; Changparallel, N. T.; Gallo, R. C.; Wong-Staal, F. *Nature* **1985**, *313*, 277–284.

(44) 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.

(45) Hyland, L. J.; Tomaszek, T. A., Jr.; Meek, T. D. *Biochemistry* **1991**, *30*, 8454–8463.

(46) Case, D.; Darden, T.; Cheatham, T., III; Simmerling, C.; Wang, J.; Duke, R.; Luo, R.; Merz, K.; Pearlman, D.; Crowley, M.; Walker, R.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D.; Schafmeister, C.; Ross, W.; Kollman, P. *AMBER 9*; University of California: San Francisco, 2006.

(47) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. *J. Comput. Chem.* **2003**, *24*, 1999–2012.

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

(49) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(50) Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269–10280.

(51) 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 02; Gaussian, Inc.: Wallingford, CT, 2004.

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

(53) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

(54) Izaguirrei, J. A.; Catarello, D. P.; Wozniak, J. M.; Skeel, R. D. *J. Chem. Phys.* **2001**, *114*, 2090–2098.

(55) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

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

(57) Laio, A.; VandeVondele, J.; Rothlisberger, U. *J. Chem. Phys.* **2002**, *116*, 6941–6947.

(58) Laio, A.; Gervasio, F. L.; VandeVondele, J.; Sulpizi, M.; Rothlisberger, U. *J. Phys. Chem. B* **2004**, *108*, 7963–7968.

(59) Becke, A. D. *Phys. Rev. A* **1998**, *38*, 3098–3100.

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

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

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

(63) Martyna, G. J.; Klein, M. L.; Tuckerman, M. *J. Chem. Phys.* **1992**, *97*, 2635–2643.

(64) Ido, E.; ping Han, H.; Kezdy, F. J.; Tang, J. *J. Biol. Chem.* **1991**, *266*, 24359–24366.

(65) Czodrowski, P.; Sotriffer, C. A.; Klebe, G. *J. Chem. Inf. Model.* **2007**, *47*, 1590–1598.

(66) Trylska, J.; Antosiewicz, J.; Geller, M.; Hodge, C.; Klabe, R.; Head, M.; Gilson, M. *Protein Sci.* **1999**, *8*, 180–195.

(67) Yamazaki, T.; Nicholson, L. K.; Wingfield, D. A. T. P.; Stahl, S. J.; Kaufman, J. D.; Eyermann, C. J.; Hodge, C. N.; Lam, P. Y. S.; Ru, Y.; Jadhav, P. K.; Chang, C.-H.; Webers, P. C. *J. Am. Chem. Soc.* **1994**, *116*, 10791–10792.

(68) Wang, Y.-X.; Freedberg, D. I.; Yamazaki, T.; Wingfield, P. T.; Stahl, S. J.; Kaufman, J. D.; Kiso, Y.; Torchia, D. A. *Biochemistry* **1996**, *35*, 9945–9950.

(69) Piana, S.; Sebastiani, D.; Carloni, P.; Parrinello, M. *J. Am. Chem. Soc.* **2001**, *123*, 8730–8737.

(70) Piana, S.; Carloni, P. *Proteins* **2000**, *39*, 26–36.

(71) Harte, W. E., Jr.; Beveridge, D. L. *J. Am. Chem. Soc.* **1993**, *115*, 3883–3886.

(72) Adachi, M.; Ohhara, T.; Kurihara, K.; Tamada, T.; Honjo, E.; Okazaki, N.; Arai, S.; Shoyama, Y.; Kimura, K.; Matsumura, H.; Sugiyama, S.; Adachi, H.; Takano, K.; Mori, Y.; Hidaka, K.; Kimura, T.; Hayashi, Y.; Kiso, Y.; Kuroki, R. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 4641–4646.

(73) Tawa, G. J.; Topol, I. A.; Burt, S. K.; Erickson, J. W. *J. Am. Chem. Soc.* **1998**, *120*, 8856–8863.

(74) Pillai, B.; Kannan, K.; Hosur, M. V. *Proteins* **2001**, *43*, 57–64.

(75) Kumar, M.; Hosur, M. V. *Eur. J. Biochem.* **2003**, *270*, 1231–1239.

(76) Davies, D. R. *Annu. Rev. Biophys. Biophys. Chem.* **1990**, *19*, 189–215.

(77) Todd, M. J.; Semo, N.; Freire, E. *J. Mol. Biol.* **1998**, *283*, 475–488.

(78) Vondrasek, J.; Wlodawer, A. *Proteins* **2002**, *49*, 429–431.

(79) Swain, A. L.; Miller, M. M.; Green, J.; Rich, D. H.; Schneider, J.; Kent, S. B. H.; Wlodawer, A. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 8805–8809.

(80) Jaskolski, M.; Tomasselli, A. G.; Sawyer, T. K.; Staples, D. G.; Heinrikson, R. L.; Schneider, J.; Kent, S. B. H.; Wlodawer, A. *Biochemistry* **1991**, *30*, 1600–1609.

(81) Barillari, C.; Taylor, J.; Viner, R.; Essex, J. W. *J. Am. Chem. Soc.* **2007**, *129*, 2577–2587.

(82) Dunitz, J. D. *Science* **1994**, *264*, 670–671.

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

(84) Schechter, I.; Berger, A. *Biochem. Biophys. Res. Commun.* **1967**, *27*, 157–162.

CT9004728

JCTC Journal of Chemical Theory and Computation

# Understanding Energetic Origins of Product Specificity of SET8 from QM/MM Free Energy Simulations: What Causes the Stop of Methyl Addition during Histone Lysine Methylation?

Yuzhuo Chu,[†] Qin Xu,[†] and Hong Guo*,[†,‡]

*Department of Biochemistry and Cellular and Molecular Biology, University of Tennessee, Knoxville, Tennessee 37996 and UT/ORNL Center for Molecular Biophysics, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-6164*

Received December 2, 2009

**Abstract:** Biological consequences of histone lysine methylation depend on the methylation states of the lysine residues on the tails of histone proteins that are methylated by protein lysine methyltransferases (PKMTs). Therefore, the ability of PKMTs to direct specific degrees of methylation (i.e., product specificity) is an important property for regulation of chromatin structure and gene expression. Here, the free energy simulations based on quantum mechanical/molecular mechanical (QM/MM) potentials are performed for the first, second, and third methyl transfers from S-adenosyl-L-methionine to the $\varepsilon$-amino group of the target lysine/methyl lysine in SET8, one of the important PKMTs. The key questions addressed in this paper include the energetic origin of the product specificity and the reasons for the change of the product specificity as a result of the replacement of Tyr334 by Phe. The free energy barriers for the three methyl transfers in SET8 as well as in the mutant obtained from the simulations are found to be well correlated with the experimental observations on the product specificity of SET8 and the change of product specificity as a result of the mutation. The results support the suggestion that the differential free energy barriers for the methyl transfers may determine, at least in part, how the epigenetic marks of lysine methylation are written by the enzymes. Furthermore, the stability of a water molecule to be located at the active site is examined under different conditions using the free energy simulations, and its role in controlling the product specificity is discussed. The QM/MM molecular dynamics (MD) simulations are also performed on the reactant complexes of the first, second, and third methyl transfers. The results show that the information on the ability of the reactant complexes to form the reactive configurations for the methyl transfers may be used as useful indicators in the prediction of product specificity for PKMTs.

## Introduction

The tails of core histone proteins in the nucleosome are subject to a variety of post-translational covalent modifications, and these modifications can be read by other proteins to lead to distinct downstream events in the regulation of chromatin structure and gene expression.[1] Protein lysine methyltransferases (PKMTs) catalyze one such modification, i.e., histone lysine methylation. Histone lysine methylation can govern a number of important biological processes, including heterochromatin formation, X-chromosome inactivation, and transcriptional silencing and activation.[2] Several lysine residues on histone proteins have been identified to be the sites of methylation, including histone H3 Lysine 4 (H3−K4), H3−K9, H3−K27, H3−K36, H3−K79, and H4−K20. In addition to selecting different lysine sites for methylation (i.e., substrate specificity), PKMTs may also

* Corresponding author e-mail: hguo1@utk.edu.

[†] University of Tennessee.

[‡] Oak Ridge National Laboratory.

10.1021/ct9006458 © 2010 American Chemical Society
Published on Web 03/17/2010

differ in their ability to transfer one, two, or three methyl groups from S-adenosyl-L-methionine (AdoMet, the methyl donor) to the $\varepsilon$-amino group of the target lysine.[3] This property of the enzymes is called product specificity. Since biological consequences of histone lysine methylation depend on the number of methyl groups added to the lysine residues, understanding the energetic origins of product specificity and developing suitable strategies for manipulation of the signaling properties are of considerable interest.

Biochemical studies and structural analyses of the SET-domain PKMT complexes have identified a tyrosine/phenylalanine switch that can control the product specificity. Comparison of the active site structure of DIM-5 (a trimethylase)[4] with that of SET7/9[5] or SET8[6] (a monomethylase) showed that a single amino acid residue occupies a structurally similar position in the enzymes (e.g., F281 in DIM-5 and Y334 in SET8) and is in proximity of the $\varepsilon$-amino group of the target lysine.[4b,7] It has been demonstrated that DIM-5 can be converted from a K-9 trimethyltransferase to a K-9 mono/dimethyltransferase by the F281→Y mutation[4b] and that the SET7/9 Y305F[4b] or SET8 Y334F mutant[5a,7] was able to generate a dimethylated instead of a monomethylated lysine product. In each case, the substrate specificity was not changed, and the mutation had little effect on the overall reaction rate. Understanding the role of the residue at this tyrosine/phenylalanine switch position and the energetic origin for the change of the product specificity as a result of the mutation can provide important insights into the property of the PKMT product specificity. Systematic determinations of the structures for SET8 and its Y334F mutant complexed with an unmodified, monomethylated, or dimethylated H4−K20 peptide along with AdoHcy were undertaken[7] to pinpoint the structural origin for the existence of the tyrosine/phenylalanine switch. It was proposed on the basis of the existence and absence of an active-site water molecule in these structures along with some other information that the Phe/Tyr switch may regulate product specificity through altering the affinity of the observed water molecule and that the dissociation of this water molecule is likely to be essential for the multiple methylation process to proceed.[7] Nevertheless, the energetic interpretations for the product specificity of SET8 and its alternation due to the Y334→F mutation are still lacking, and it is not clear as to how the enzyme's ability to stabilize/destabilize the water molecule at the active site would change as a result of the mutation and/or methyl addition.

Computer simulations can provide important insights into the energetic origins of the product specificity as well as the effects of mutations. One approach that has been used previously for PKMTs is to perform the free energy (potential of mean force) simulations with hybrid quantum mechanical/molecular mechanical (QM/MM) potentials and to establish the correlations between the free energy profiles of the methyl transfers and the product specificity[8] (see below for the discussions of the results based on other computational approaches). In our earlier communication,[8b] it was demonstrated that, for DIM5, SET7/9, and their certain mutants, the three free energy barriers for the methyl transfers may be used in each case to explain the product specificity observed experimentally. It was hypothesized[8b] that the relative efficiencies of the chemical steps involving the three methyl transfers from AdoMet to the $\varepsilon$-amino group of the target lysine in PKMTs may determine, at least in part, the product specificity. The results of the QM/MM molecular dynamics (MD) simulations on the reactant complexes have also been compared for the first, second, and third methyl transfers for SET7/9, DIM-5, and their mutants.[8b] It was shown that a correlation may be established between the formation of the reactive configurations for the three methyl transfers and the product specificities of the enzymes. One problem in the earlier work[8b] is that the experimental structures for the SET7/9 and DIM-5 complexes with *different methylation states* for the target lysine residues do not exist, and some manual modifications had to be introduced in the generation of suitable reactant complexes containing methylated lysine. In order to establish the prediction on the relationship between the efficiency of the methyl transfers and product specificity, additional simulations need to be performed on the PKMT (and the mutant) complexes for which the X-ray structures have been determined at different methylation states for the target lysine residue. SET8 is an excellent system for such investigations because of the recent availability of several experimental structures with unmodified, mono- and dimethylated lysine residue at the active site. Furthermore, the location of the important active-site water molecule has also been clearly identified in the X-ray structures. The simulations based on these structures may not only lead to a better understanding of the energetic origin of the product specificity but also provide important energetic information concerning the stability of this water molecule at the active site at different stages of methylation in different systems that is believed to be a key property of the enzymes in controlling the product specificity.[7]

Here, we report the results of QM/MM free energy simulations on SET8 and its Y334F mutant. The free energy barriers for the methyl transfers in SET8 and the mutant obtained from the simulations are found to be well correlated with the experimental observations on the product specificities, supporting the suggestion that the differential free energy barriers for the methyl transfers may determine, at least in part, how the epigenetic marks of lysine methylation are written by the enzymes. Furthermore, the stability of the water molecule at the active site under different conditions (see above) is also examined on the basis of the free energy simulations. The free energy profiles show that the stability of the water molecule at the active site decreases significantly as a result of the Y334→F mutation as well as the methyl addition to the lysine residue. Such changes are likely to make it easier for the water molecule to dissociate from the active site and create the space for further methyl addition. The QM/MM MD simulations are also performed on the reactant complexes of the first, second, and third methyl transfers. The results show that the dynamic information on the ability of the reactant complexes to form the reactive configurations for the methyl transfers may be used as useful indicators in the prediction of product specificity for PKMTs, although other factors can be involved as well.

## Methods

QM/MM free energy (potential of mean force) and MD simulations were applied to determine free energy profiles for the first, second, and third methyl transfers from AdoMet to the $\varepsilon$-amino group of the target lysine (methyl lysine) and to characterize the active-site dynamics of the reactant complexes of the methyl transfers in SET8 and the Y334F mutant using the CHARMM program.[9] AdoMet/AdoHcy and lysine/methyl-lysine side chains were treated by QM and the rest of the system by MM. The link-atom approach[10] as implemented in CHARMM was applied to separate the QM and MM regions. Although the QM/MM approach in principle is not required for MD investigations of the reactant complexes, the previous studies on SET7/9 and DIM-5[8] showed that the QM/MM MD approach seems to provide a good description of the active-site dynamic features of the reactant complexes that are consistent with experimental observations concerning the product specificity. A modified TIP3P water model[11] was employed for the solvent. The stochastic boundary molecular dynamics method[12] was used for the QM/MM MD and free energy simulations. The system was separated into a reaction zone and a reservoir region, and the reaction zone was further divided into a reaction region and a buffer region. The reaction region was a sphere with radius $r$ of 20 Å, and the buffer region extended over 20 Å $\leq r \leq$ 22 Å. The reference center for partitioning the system was chosen to be the $C_\delta$ atom of the target lysine residue/methyl lysine. The resulting systems contained around 5500 atoms, including about 800−900 water molecules.

The SCC-DFTB method[13] implemented in CHARMM was used for the QM atoms, and the all-hydrogen CHARMM potential function (PARAM27)[14] was used for the MM atoms. High-level ab initio methods (e.g., B3LYP and MP2) are too time-consuming to be used for MD and free energy simulations. The results of the SCC-DFTB and B3LYP/6-31G** methods for the description of the methyl transfer in a small model system have been compared in earlier studies[8] using an energy-minimization-based approach. This comparison allowed us to understand the performance of the semiempirical method in the description of the bond breaking and making for the system under investigation. It was found that, although the SCC-DFTB optimized geometries along the reaction pathway seemed to be rather close to those from B3LYP/6-31G**, there are some systematic deviations of the SCC-DFTB method in the description of the energetics of the methyl transfer. To correct the errors due to the deficiency of the SCC-DFTB method, the empirical correction introduced in the earlier studies[8] was applied to the free energy curves obtained from the potential of mean force simulations in the present work (see below). In the earlier study of DIM-5 (see the Supporting Information in ref 8b), it was shown that that the energy curves from the corrected SCC-DFTB and B3LYP/6-31G** were very close, supporting the use of the approach with the empirical correction. It should be pointed out that the reason that the simple empirical correction can be used in this and previous studies is because the bond breaking and making events in this and the previous papers all involve simple and similar $S_N2$ methyl transfer processes so that most of the errors are expected to be canceled out. Moreover, the relative free energy barriers, as opposed to the absolute barriers, are expected to be more important in the determination of the product specificity. The relative free energy barriers are expected to be less sensitive to the choice of the QM method due to the cancellation of the errors. Indeed, the results of our earlier simulations[8a] were confirmed by the use of a quite different QM/MM approach[15] with a difference of only about 1 kcal/mol in the relative free energy barriers of the first and second methyl transfers in SET7/9.

The initial coordinates for the reactant complexes of the first, second, and third methyl transfers were based on the crystallographic complexes (PDB codes: 1ZKK, 3F9W, 3F9X, and 3F9Y) of SET8 and its Y334F mutant containing AdoHcy and short H4K20, H4K20me1, and H4K20me2 peptides.[5a,7] In all the cases, a methyl group was manually added to AdoHcy to form AdoMet. In addition, for the reactant complex of the second methyl transfer in the wild type, a methyl group was manually added to the target lysine in the X-ray structure (1ZKK) to form the methylated lysine. The initial structures for the entire stochastic boundary systems were optimized using the steepest descent (SD) and adopted-basis Newton−Raphson (ABNR) methods. The systems were gradually heated from 50.0 to 310.15 K over 50 ps. A 1 fs time step was used for integration of the equation of motion, and the coordinates were saved every 50 fs for analyses. The 1.5 ns QM/MM MD simulations were carried out for each of the reactant complexes of the first, second, and third methyl transfers, and the data from the final 0.5 ns were used to generate the distribution maps of $r(C_M-N_\xi)$ and $\theta$ in each case (see below). As discussed in the previous studies and mentioned earlier in this paper, the $S_N2$ methyl transfer from AdoMet to H4−K20, H4−K20me1, or H4−K20me2 is presumably more efficient if the $S-CH_3$ group of AdoMet is well aligned with the lone pair of electrons on $N_\xi$ in the reactant complex, i.e., with a small $\theta$ angle and relatively short $C_M-N_\xi$ distance.[8] Here, $\theta$ is defined as the angle between the direction of the $C_M-S_\delta$ bond ($r_2$) and the direction of the electron lone pair ($r_1$) (see Figure 1). Therefore, we determined the distributions of $r(C_M-N_\xi)$ and $\theta$ from the QM/MM MD trajectories to obtain the information about the relationship between these distributions and product specificity. Moreover, the histogram method was used to calculate the probability density distributions of $r(C_M-N_\xi)$ and $\theta$ and the relative free energies as functions of $r(C_M-N_\xi)$ and $\theta$. For $r(C_M-N_\xi)$, histograms with a bin width of 0.1 Å were used, and the probability density in the $i$th histogram is as follows: $\rho_i = N_i/N$ ($N$ is the total number of the configurations from the MD simulations, and $N_i$ is the occurrence number in the $i$th histogram). To calculate the probability density distribution of $\theta$, the histograms with a width of 10° were used, and $N_i$ was weighted by $1/A_i$, where $A_i$ is the area of the $i$th histogram of $\theta$. Thus, the probability density in the $i$th histogram of $\theta$ is as follows: $N_i/(N \times A_i)$. The relative free energy of the $i$th histogram of $r(C_M-N_\xi)$ and $\theta$ were calculated through $W_i =$

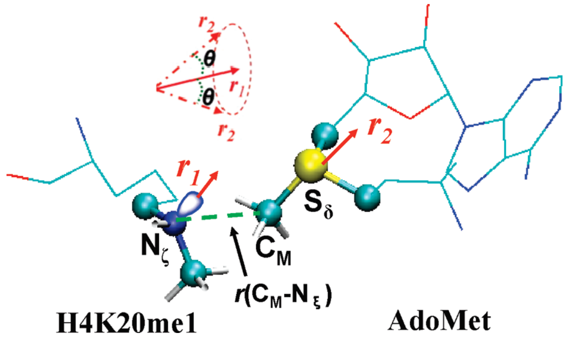


***Figure 1.*** Definition of the structural parameters for monitoring the relative orientation of AdoMet and H4K20me1 [H4K20 and H4K20(me)$_2$] in the reactant complex. The efficiency of the methyl transfer may be related to the distributions of $r(C_M\cdots N_\zeta)$ and $\theta$ in the reactant complexes. $\theta$ is defined as the angle between the two vectors $r_1$ and $r_2$. Here, $r_1$ is the direction of the lone pair of electrons on $N_\zeta$ and $r_2$ is the vector pointing from $C_M$ to $S_\delta$. The reaction coordinate for calculating the free energy profiles for the methyl transfers is $R = r(C_M\cdots S_\delta) - r(C_M\cdots N_\zeta)$.

$-k_BT \times \ln \rho_i$, where $k_B$ is Boltzmann's constant and $T$ is the temperature (see the Supporting Information of ref 8a for additional information).

The umbrella sampling method[16] implemented in the CHARMM program along with the Weighted Histogram Analysis Method (WHAM)[17] was applied to determine the change of the free energy (potential of mean force) as a function of the reaction coordinate for the methyl transfer from AdoMet to H4−K20, H4−K20me1, or H4−K9me2 in the wild-type and mutated enzymes. The reaction coordinate was defined as a linear combination of $r(C_M-N_\xi)$ and $r(C_M-S_\delta)$ [$R = r(C_M-S_\delta) - r(C_M-N_\xi)$] (see Figure 1). For each methyl transfer process, 20 windows were used, and for each window, 50 ps production runs were performed after 20 ps equilibration. The force constants of the harmonic biasing potentials used in the PMF simulations were 50 to 500 kcal mol$^{-1}$ Å$^{-2}$. The statistical errors for the free energy profiles were also estimated and were found to be quite small (see the Supporting Information). In addition, the umbrella sampling method was applied to determine the free energy profiles for the movement of the active-site water molecule (W1) in the wild type and Y334F at different methylation states. The reaction coordinate was defined as the distance between the water oxygen and sulfur atom of AdoMet [i.e., $r(O_{w1}\cdots S_\delta)$]. Twenty windows were used for the change of the water position, and for each window, 20 ps production runs were performed after a 20 ps equilibration. The force constants of the harmonic biasing potentials used in the PMF simulations were 20 to 50 kcal mol$^{-1}$ Å$^{-2}$. The distance for which the simulations were performed is in the range of 3−15 Å. The simulations were not performed for longer distances, because the use of the stochastic boundary method in the present study may limit the flexibility of the water molecule as it moves closer to the boundaries of the models (at ~22−25 Å from sulfur atom). Additional simulations with different boundary conditions will be performed in the future.

## Results

The average active-site structure of the reactant complex for the first methyl transfer in SET8 is given in Figure 2A. Figure 2A shows that the active-site structure has the lone pair of electrons on $N_\zeta$ of the target lysine well aligned with the methyl group of AdoMet. This is further demonstrated by the large population of the structures with relatively short $r(C_M\cdots N_\zeta)$ distances and small values of the $\theta$ angle as well as the free energy plots generated from the results of the simulations (Figure 2B). The average distance between $N_\xi$ and the methyl group ($C_MH_3$) is approximately 3.0 Å, and the angle is mainly in the range of 0−30°. Figure 2A also shows that Tyr245 forms a hydrogen bond with the ε-amino group of the target lysine, and this hydrogen bond may help to orientate the direction of the electron lone pair toward the methyl group of AdoMet. A water molecule (W1) forms stable hydrogen bonds with the both ε-amino groups of H4K20 and Tyr334 (the important tyrosine/phenylalanine switch residue, see above). Figure 2D and E show that, for the reactant complex of the second methyl transfer, the average distance between $N_\xi$ and the methyl group (~4.5 Å) and the values of $\theta$ (mainly in the range of 45−120°) become significantly larger compared to those for the first methyl transfer (Figure 2A and B). Thus, the S−CH$_3$ group of AdoMet cannot be well aligned with the lone pair of electrons on $N_\xi$ for the second methyl transfer, suggesting that the efficiency of the corresponding methyl transfer may be significantly compromised. Indeed, Figure 2E shows that the free energy cost for producing a structure like the one in Figure 2A is approximately 4−5 kcal/mol [i.e., ~3 kcal/mol for changing $r(C_M\cdots N_\zeta)$ from 4.5 Å to 3 Å and 1.5 kcal/mol for changing $\theta$ to less than 30°].

The average structures for the first, second, and third methyl transfers in Y334F are given in Figure 3. As is evident from Figure 3C, the lone pair of electrons on $N_\zeta$ of the methyl lysine is well aligned with the methyl group of AdoMet for the second methyl transfer in Y334F. This is in contrast to the case for the second methyl transfer in the wild type for which the two cannot be well aligned (see above). The results suggest that the efficiency of the second methyl transfer is likely to be significantly enhanced due to the improvement of the reactant structure for the methyl transfer, although other factors may be involved as well (see below). For the third methyl transfer in Y334F (Figure 3E), the S−CH$_3$ group of AdoMet cannot be well aligned with the lone pair of electrons on $N_\xi$ in the reactant complex as indicated by the long $r(C_M\cdots N_\zeta)$ distance (~4.6 Å) and large values of the $\theta$ angle (45−150°). Thus, the corresponding methyl transfer is unlikely to be efficient. One of the key structural changes at the active site is the relocation of the active-site water molecule (W1). Indeed, W1 occupies a position in Figure 3C that is completely different from the one in Figure 2D (i.e., the complex for the second methyl transfer in the wild type) and is not in close contact with the target lysine/methyl lysine and the Y334/F334 residue anymore. For the third

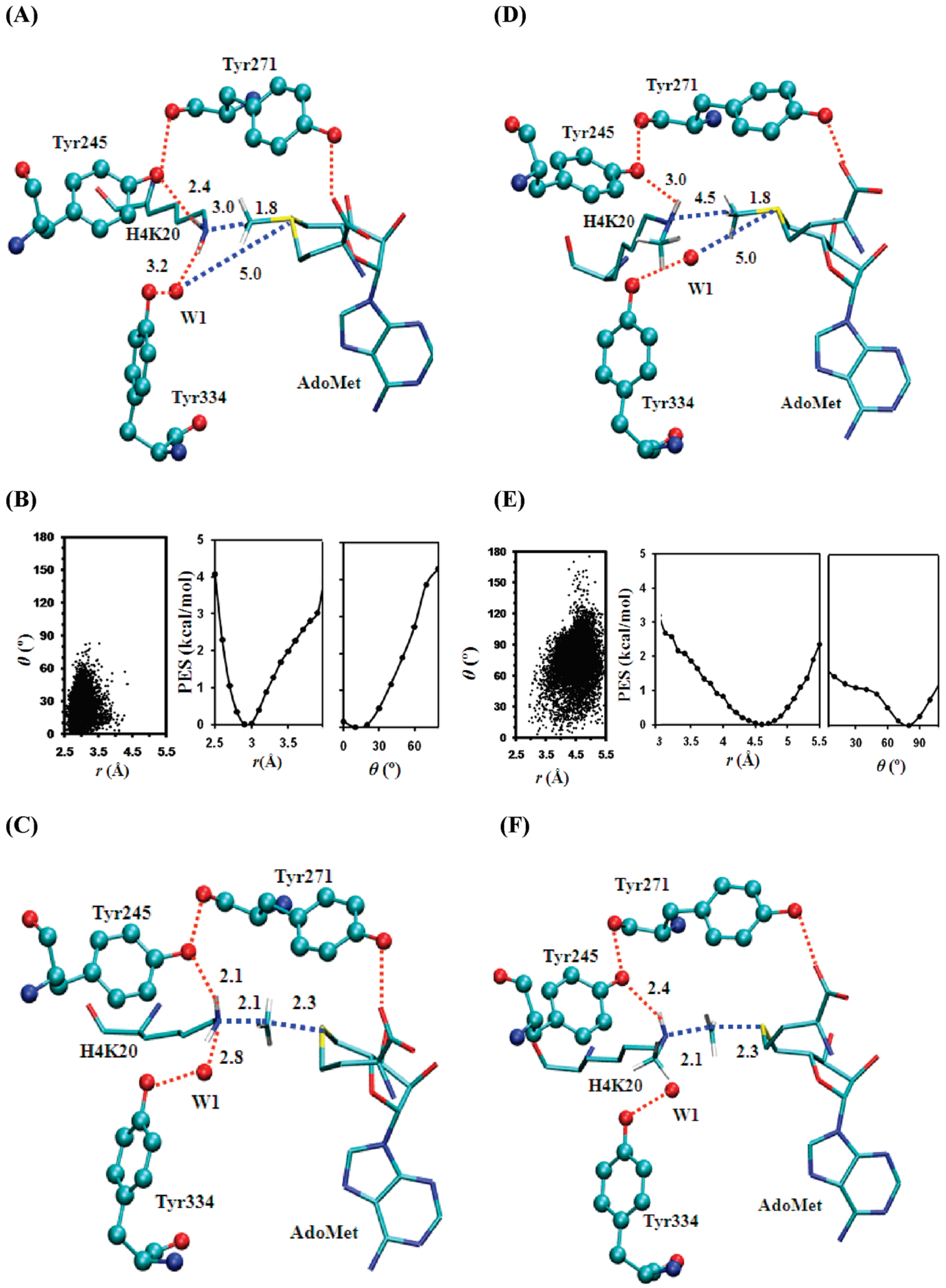


***Figure 2.*** MD results for the wild-type enzyme (SET8). (A) The average active-site structure of the reactant complex for the first methyl transfer. SET8 is shown in balls and sticks, and AdoMet and the H4K20 side chain are in sticks. Hydrogen atoms are not shown for clarity, except for those on $N_\zeta$ and the transferable methyl group. Hydrogen bonds are indicated by red dotted lines, and the distances related to the reaction coordinates are also shown. (B) Left: the two-dimensional plot of $r(C_M \cdots N_\zeta)$ and $\theta$ distributions based on the 1.5 ns simulations of the reactant complex for the first methyl transfer. Middle: the free-energy change as a function of $r(C_M \cdots N_\zeta)$ obtained from the distributions. Right: the free-energy change as a function of $\theta$ obtained from the distributions. (C) The average structure near the transition state for the first methyl transfer obtained from the free energy (potential of mean force) simulations. (D) The average structure of the reactant complex for the second methyl transfer. (E) Left: the two-dimensional plot of $r(C_M \cdots N_\xi)$ and $\theta$ distributions of the reactant complex for the second methyl transfer. Middle: the free-energy change as a function of $r(C_M \cdots N_\xi)$ obtained from the distributions. Right: the free-energy change as a function of $\theta$ obtained from the distributions. (F) The average structure near the transition state for the second methyl transfer.

methyl transfer in Y334F, W1 has been pushed away from the active site during the MD simulations and is not visible in Figure 3E.

The free-energy profiles for the first and second methyl transfers in SET8 are plotted in Figure 4A as a function of the reaction coordinate; the free energy barrier for the first methyl transfer was calculated to be 15.8 kcal/mol. This free energy barrier is within the error limit of the average barrier from earlier single-point MP2/6-31G+G(d,p)/MM calculations (13.9 ± 2.3 kcal/mol) on SET8.[18] It should be pointed out, however, that cautions must be exercised when the energetic data obtained on the basis of very different computational approaches are compared. As is evident from Figure 4A, the free energy barrier for the second methyl

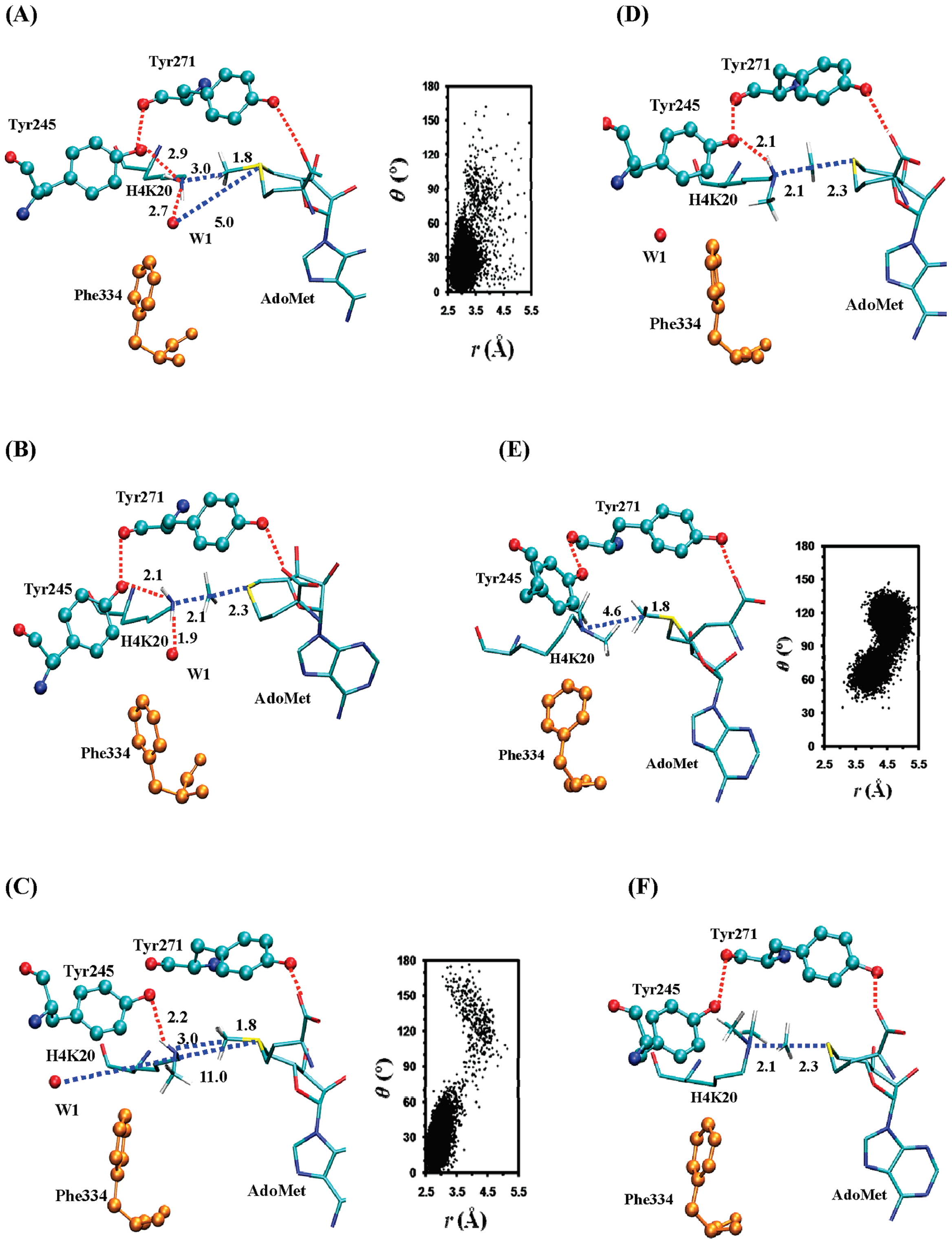


***Figure 3.*** MD results for Y334F. (A) Left: the average active-site structure of the reactant complex for the first methyl transfer. Right: the two-dimensional plot of $r(C_M \cdots N_\zeta)$ and $\theta$ distributions of the reactant complex for the first methyl transfer. (B) The average structure near the transition state for the first methyl transfer. (C) Left: the average structure of the reactant complex for the second methyl transfer. Right: the two-dimensional plot of $r(C_M \cdots N_\zeta)$ and $\theta$ distributions of the reactant complex for the second methyl transfer. (D) The average structure near the transition state for the second methyl transfer. (E) Left: the average structure of the reactant complex for the third methyl transfer. Right: the two-dimensional plot of $r(C_M \cdots N_\zeta)$ and $\theta$ distributions of the reactant complex for the third methyl transfer. (F) The average structure near the transition state for the third methyl transfer.

transfer is much higher than that of the first methyl transfer (by as much as 6.5 kcal/mol). Thus, the second methyl transfer is much less efficient compared to the first methyl transfer process, and this is consistent with the experimental findings that SET8 is a monomethylase.[5,7] The earlier single-point MP2/6-31G+G(d,p)/MM calculations led to an average

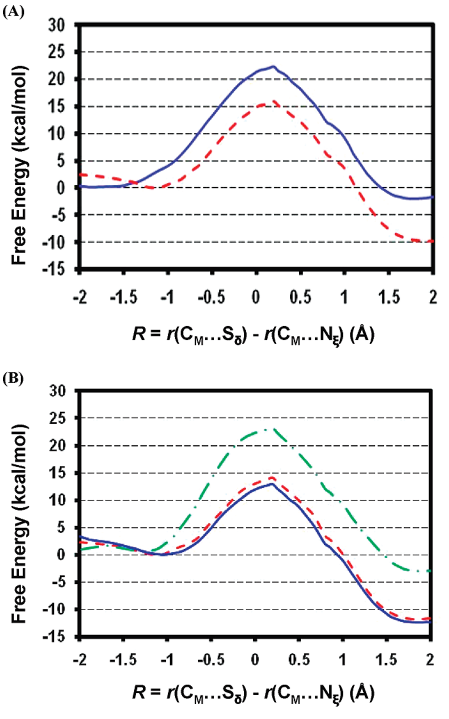


***Figure 4.*** (A) Free energy (potential of mean force) changes for the first and second transfers from AdoMet to H4−K20 and H4−K20me1, respectively, as a function of the reaction coordinate [$R = r(C_M\cdots S_\delta) - r(C_M\cdots N_\xi)$] in the wild-type SET8. The first methyl transfer: red and dashed line with a free energy barrier of 15.8 kcal/mol. The second methyl transfer: blue and solid line with a free energy barrier of 22.3 kcal/mol (or about 6.5 kcal/mol higher than that of the first methyl transfer). Differences in the free energy barriers may be represented by two energy triplets, ($0, \Delta_{2-1W}, \Delta_{3-1W}$) and ($\Delta_{M-W}, \Delta_{2-1M}, \Delta_{3-1M}$), for the wild-type and mutated enzymes, respectively. For the wild-type enzyme, the second ($\Delta_{2-1W}$) and third ($\Delta_{3-1W}$) parameters are the differences in the free energy barriers between the second and first and between the third and first methyl transfers, respectively. For the mutated enzyme, the first parameter ($\Delta_{M-W}$) is the difference in the free energy barriers for the first methyl transfer in the wild-type and mutant. The second ($\Delta_{2-1M}$) and third ($\Delta_{3-1M}$) parameters are the differences in the free energy barriers between the second and first and between the third and first methyl transfers, respectively, in the mutant. For SET8, ($0, \Delta_{2-1W}, \Delta_{3-1W}$) = ($0, 6.5, x$) ($x$ indicates the undetermined relative barrier in the energy triplet). (B) The free energy changes for the first, second, and third methyl transfers as a function of the reaction coordinate in the Y334F mutant. The first methyl transfer: red and dashed line with a free energy barrier of 14.1 kcal/mol. The second methyl transfer: blue and solid line with a free energy barrier of 13 kcal/mol. The third methyl transfer: green and dot-dashed line with a free energy barrier of 23.1 kcal/mol. ($\Delta_{M-W}, \Delta_{2-1M}, \Delta_{3-1M}$) = (−1.7, −1.1, 9).

barrier that is as much as 20 kcal/mol higher for the second methyl transfer than that for the first methyl transfer.[18] The ab initio QM [HF(6-31G*/3-21G*)]/MM free energy simulations for the first and second methyl transfers in two different PKMTs, SET7/9 and Rubisco LSMT, have also been performed previously.[15a] For SET7/9, the free energy barriers for the first and second methyl transfers were calculated to be 22.5 ± 0.5 kcal/mol and 26.2 ± 0.5 kcal/mol, respectively.[15a] Comparing with our earlier results on SET7/9[8a] and taking into account the suggestion that the HF(6-31G*/3-21G*) method may overestimate the barriers by about 3 kcal/mol,[15a] these data indicate that the corrected SCC-DFTB method might underestimate the barriers, although other factors may affect the results as well. However, as mentioned earlier, the relative free energy barriers, instead of the absolute barriers, are expected to be more important in the determination of the product specificity. For SET7/9, the difference for the two methods in the description of the differential free energy barriers is only about 1 kcal/mol.

Figure 4B plots the free energy profiles for the methyl transfers in the Y334F mutant. Unlike the wild-type enzyme, the free energy profiles for both the first and second methyl transfers are rather similar with relatively low barriers. Thus, if the first methyl transfer from AdoMet to the target lysine can be catalyzed by Y334F, the second methyl transfer to monomethyl lysine would also be possible. By contrast, the free energy barrier for the third methyl transfer in Y334F is considerably higher. The results are consistent with the experimental observations that Y334F is a dimethylase[5,7] and support the suggestion[8] that the relative free energy barriers for the methyl transfers are likely to be important energetic factors controlling the product specificity.

Figure 5A and B plot the free energy profiles as a function of the distance between the oxygen atom of W1 and the sulfur atom of AdoMet in the reactant complexes for the first and second methyl transfers, respectively, in the wild-type enzyme. As is evident from Figure 5A and B, the most stable location for W1 is at the active site in these complexes, with a distance of about 5 Å to the sulfur atom, consistent with the average structures observed from the MD simulations (Figure 2A and D). The free energy profiles further show that W1 seems to be held tightly by the active site interactions, as the energetic cost for W1 to move away from the active site in each case is quite high. Figure 5D shows that, for the reactant complex of the second methyl transfer in Y334F, the most stable position for W1 has moved away from the active site (to a location with a distance of about 10−11 Å from the sulfur atom). With the removal of W1, the active site of Y334F becomes less crowded and is able to accommodate the second methyl group on the target lysine. The second methyl transfer can therefore proceed, and the mutant becomes dimethylase (see below).

## Discussions

The key question on the product specificity of PKMTs concerns the factor that controls *the methylation state of the product*. This is in contrast with many other investigations on enzyme-catalyzed reactions which concentrate on the effects of enzymes in the reduction of the activation barriers in going from solution to the enzyme active sites. Thus, a convenient reference reaction for understanding the product specificity would be the process involving the first methyl transfer in the wild-type enzyme (see below). The existence of a relatively high barrier for one of the methyl transfer processes may lead to the termination of further methyl addition and therefore determine the product specificity of the enzyme. If none of the three free-energy barriers for the

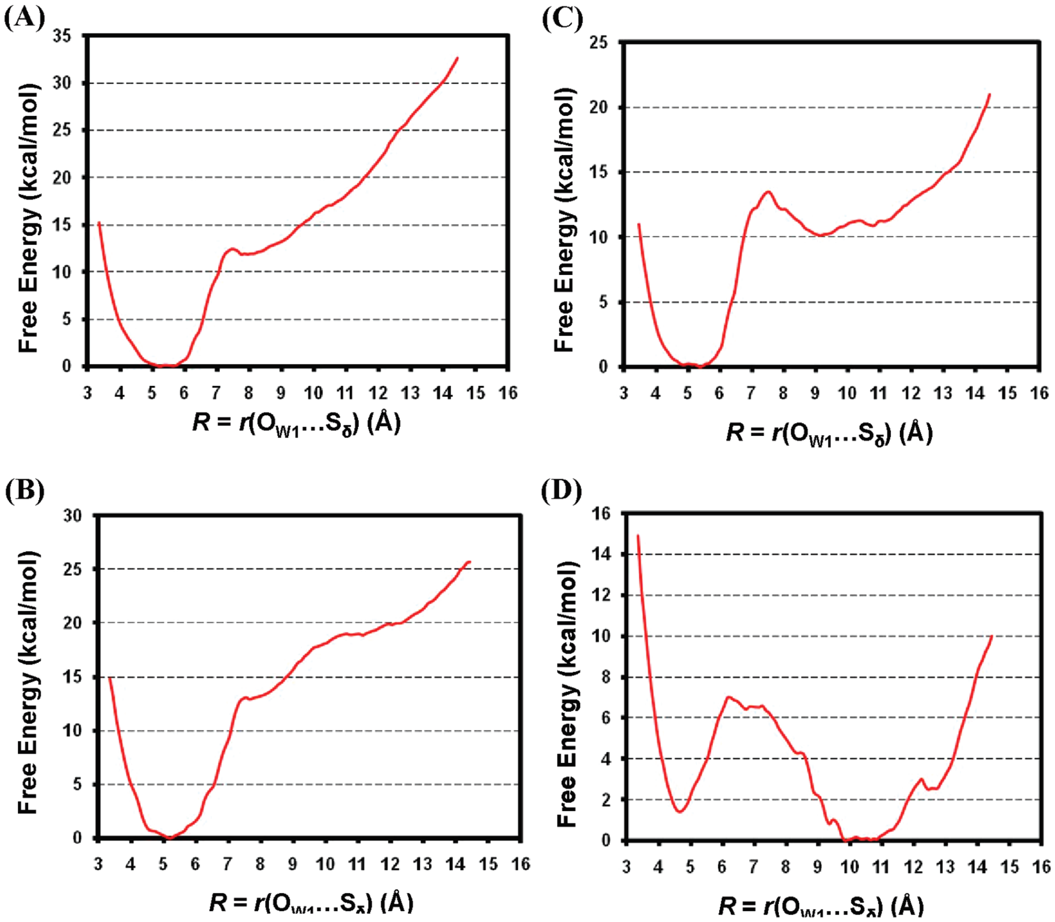


***Figure 5.*** (A) Free energy (potential of mean force) change as a function of the distance between the oxygen atom of the active-site water molecule (W1) and sulfur atom of AdoMet [i.e., $r(O_{w1}\cdots S_\delta)$] in the reactant complex for the first methyl transfer in the wild-type enzyme. W1 is stable in the active site, as the free energy minimum is around 5 Å (i.e., at a similar position to that observed in the MD simulations in Figure 2A). High energy at the longer $r(O_{w1}\cdots S_\delta)$ distance suggests that it would be difficult for W1 to move away from the active site. (B) Free energy change in the reactant complex of the second methyl transfer in the wild-type enzyme. The simulation data show that W1 is also stable in the active site. (C) Free energy change in the reactant complex of the first methyl transfer in Y334F. (D) Free energy change in the reactant complex of the second methyl transfer in Y334F. The most stable position for W1 has changed (now 10−11 Å from the sulfur atom), suggesting that W1 is not in the active site.

methyl transfers is significantly high, the enzyme might be able to catalyze all three methyl transfers and could be a trimethylase (e.g., as in the case of DIM-5; see ref 8b). This proposal is consistent with some previous computational results that suggested that the product specificity is probably to be mainly controlled by the methyl transfer reaction step.[8,15] An alternative explanation of the product specificity is based on the formation of a water channel observed during the MD simulations of PKMTs.[18,19] However, the dramatic increase of the energy barrier from the first to the second methyl transfer in SET8 obtained by the same authors[18] (see above) raises the question concerning the true event that prevents the further methylation in PKMTs (e.g., dimethylation by SET8).

It was proposed in our earlier communication that two different free energy triplets, (0, $\Delta_{2-1W}$, $\Delta_{3-1W}$) and ($\Delta_{M-W}$, $\Delta_{2-1M}$, $\Delta_{3-1M}$) for wild-type and mutated enzymes, respectively, may be used in the description of the product specificity of PKMTs and their mutants.[8b] Here, the free energy barrier for the first methyl transfer in the wild-type enzyme (i.e., the reference reaction) is taken as the zero (for a detailed explanation of the parameters, see Figure 4). For SET8 and its Y334F mutant studied in this work, the corresponding energy triplets can be written as (0, 6.5, $x$) and (−1.7, −1.1, 9), respectively, which reflect the fact that SET8 is a monomethylase and the mutant a dimethylase (see below). The kinetic study on Y334F[7] suggested that the activation barrier to produce the H4K20me2 product from the H4K20me1 substrate is about 2 kcal/mol higher than the barrier to produce the monomethyl product from the unmodified H4K20 substrate. Figure 4B shows that the second methyl transfer has a free energy barrier that is slightly lower than that for the first methyl transfer (by ~1 kcal/mol). The simulation data are therefore consistent with the suggestion[7] that methyl lysine reorientation and deprotonation between turnovers may constitute a rate-limiting step in catalysis. However, it should be pointed out that, as far as the product specificity is concerned, the key question is what causes the stop of further methyl addition during histone lysine methylation (which is different from the question concerning the rate-limiting step of the enzyme-catalyzed process). The results of the simulations from the present work on SET8 and our previous studies on DIM-5 and SET7/9[8b] (and certain mutants) suggest that one of the key energetic factors for the specific product specificities of PKMTs is presumably due to a significant increase of the free energy barrier for one of the methyl transfers in the enzymes. Thus, the reason that SET8 is monomethylase is probably due to the fact that the energy barrier for the second methyl transfer is too high and stops further methylation. The Y334→F mutation on SET8 effectively reduces the barrier for the second methyl transfer so that the second methyl transfer can proceed. Since the free energy barrier for the third methyl transfer is very high, the addition of the third methyl group cannot proceed, leading to a dimethylase. Similar arguments may be made for SET7/9 and DIM-5.

It is interesting to note that the free energy data on the ability of SET8 (Y334F) to catalyze the first (first and second) methyl transfer and the inability of SET8 (Y334F) to catalyze the second (third) methyl transfer are already reflected from the MD simulations in Figure 2B and E (Figure 3A, C, E). Similar observations have also been made previously.[8,15,18] Thus, the dynamic information on the ability of the reactant complexes to form the reactive configurations for the methyl transfers may be used as useful indicators in the prediction of product specificity for PKMTs, although further tests are still necessary to establish the correlations. This result is of importance, because performing the MD simulations is much easier than undertaking the QM/MM free energy simulations. Examination of the structures at the transition states (TSs) in Figures 2 and 3 shows that all these structures are rather similar; e.g., $r(\mathrm{C_M}\cdots\mathrm{N}_\zeta)$ and $r(\mathrm{S}_\delta\cdots\mathrm{C_M})$ are 2.1 and 2.3 Å, respectively, in each of the structures. It is of interest to note these structures are rather close to the TS structures generated earlier from the ab initio QM [HF(6-31G*/3-21G*)]/MM free energy simulations for two different PKMTs, SET7/9 and LSMT.[15a] As discussed earlier, the structures for the corresponding reactant complexes, on the other hand, can be significantly different. Indeed, the structures for the reactant complexes of the first methyl transfer in the wild type (Figure 2A) and the first and second methyl transfers in Y334F (Figure 3A and C, respectively) are rather similar to the corresponding TS structures (Figures 2C and 3B and D, respectively) with a $r(\mathrm{C_M}\cdots\mathrm{S}_\delta)$ distance of about 4.8 Å. For these cases, a part of the TS stabilization is probably already reflected in the reactant state through the generation of such a TS-like conformation. By contrast, the structures of the reactant complexes for the second methyl transfer in the wild type (Figure 2D) and the third methyl transfer in Y334F (Figure 3E) are significantly distorted from the corresponding TS structure (Figures 2F and 3F, respectively), and the $r(\mathrm{C_M}\cdots\mathrm{S}_\delta)$ distances are around 6.3 Å. Therefore, additional energetic cost would be required to generate the TS-like structures from these structures, and this could lead to relatively high activation barriers for the corresponding methyl transfers. It should be pointed out that the free energy costs for generating the TS-like structures alone seem not to be sufficient to explain the increases of the barriers for the methyl transfers. Indeed, Figure 2E shows that the free energy cost for producing a structure similar to the one in Figure 2A is approximately 4−5 kcal/mol, while the free energy barrier increases by 6.5 kcal/mol. Therefore, other factors may be involved as well. It is of interest to note from Figure 2A and C that the hydrogen bond distances involving the $\varepsilon$-amino group (with Tyr245 and W1) decrease significantly as the system reaches the transition state. This indicates that the corresponding interactions are strengthening and may play an important role in the transition state stabilization for the methyl transfer as well.

It was proposed that the Phe/Tyr switch may regulate product specificity through altering the affinity of W1, and the dissociation of this water molecule is essential for the multiple methylation process to proceed.[7] Although the structural information on the presence and absence of W1 under different conditions is of considerable interest, question remains concerning how the affinity of W1 would change as a result of the mutation (or methyl addition) and what would be its stability to be located in the active site. The energetic information concerning the stability of W1 at the active site under different conditions is of fundamental importance for the determination of the role of this water molecule in preventing further methlation. Figure 5B shows that W1 is rather stable in the reactant complex of the second methyl transfer in the wild-type enzyme and the free energy cost for its removal is quite high. The high stability is presumably achieved through the interaction involving Tyr334 (Figure 2D). The high energetic cost can make the addition of the second methyl group much more difficult, presumably because the active site becomes too crowded without the removal of W1. This could contribute to the significant increase of the free energy barrier from the first to the second methyl transfer and stop the second methyl addition, although additional simulations are still necessary. Figure 5D shows that, for the reactant complex of the second methyl transfer in Y334F, the most stable location of W1 has changed and W1 moved away from the active site (see also Figure 3D). Thus, this water molecule could not interfere with the methyl transfer process anymore. The mutant is now able to catalyze the second methyl transfer and becomes a dimethlase.

## Conclusions

The QM/MM free energy simulations have been performed for the first and second methyl transfers from AdoMet to the target lysine/methyl lysine in SET8 and for the first, second, and third methyl transfers in its Y334F mutant (involving the replacement of the tyrosine/phenylalanine switch residue). The two free energy barriers for the methyl transfers in SET8 and the three free barriers in the mutant obtained from the simulations have been found to be well correlated with the experimental observations on their product specificities. The results indicated that the significant increase of the free energy barrier for the second methyl transfer in SET8 (for the third methyl transfer in Y334F) might stop further methyl addition, and this could be the reason that SET8 (Y334F) is a monomethylase (dimethylase). The results support an earlier suggestion[8b] that the differential free energy barriers for the methyl transfers may determine, at least in part, how the epigenetic marks of lysine methylation are written by the enzymes. The QM/MM molecular dynamics (MD) simulations are also performed on the reactant complexes of the first and second methyl transfers in SET8 and the first, second, and third methyl transfers in Y334F. The results showed that the dynamic information on the ability of the reactant complexes to form the reactive configurations for the methyl transfers might be used as useful indicators in the prediction of product specificity for PKMTs. The stability of the water molecule at the active site has also been examined on the basis of the free energy simulations. The free energy profiles suggested that the stability of the water molecule at the active site decreases significantly as a result of the Y334→F mutation as well as the methyl addition to the lysine residue. The decrease of the stability of W1 to be located at the active site as a result of the Y334→F mutation is likely to make it easier for the water molecule to dissociate from the active site and create space for further methyl addition.

**Acknowledgment.** We thank Dr. Hao-Bo Guo, Mr. Mauricio Valenzuela, and Professors Xiaodong Cheng and Jeremy C. Smith for useful discussions and Martin Karplus for a gift of the CHARMM program. This work was supported by the National Science Foundation (Grant number: 0817940 to H.G.). We are also grateful for the computer time from Indiana University TeraGrid Resource.

**Supporting Information Available:** The estimates of the free energies for formation of the reactive conformations for Y334F and the statistical errors in the potential of mean force simulations. This information is available free of charge via the Internet at http://pubs.acs.org/.

## References

(1) (a) Taverna, S. D.; Li, H.; Ruthenburg, A. J.; Allis, C. D.; Patel, D. J. *Nat. Struct. Mol. Biol.* **2007**, *14*, 1025–1040. (b) Lall, S. *Nat. Struct. Mol. Biol.* **2007**, *14*, 1110–1115. (c) Turner, B. M. *Nat. Struct. Mol. Biol.* **2005**, *12*, 110–2.

(2) (a) Jenuwein, T. *FEBS J.* **2006**, *273*, 3121–35. (b) Martin, C.; Zhang, Y. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 838–49.

(3) (a) Xiao, B.; Wilson, J. R.; Gamblin, S. J. *Curr. Opin. Struct. Biol.* **2003**, *13*, 699–705. (b) Cheng, X.; Collins, R. E.; Zhang, X. *Annu. Rev. Biophys. Biomol. Struct.* **2005**, *34*, 267–94.

(4) (a) Zhang, X.; Tamaru, H.; Khan, S. I.; Horton, J. R.; Keefe, L. J.; Selker, E. U.; Cheng, X. *Cell* **2002**, *111*, 117–27. (b) Zhang, X.; Yang, Z.; Khan, S. I.; Horton, J. R.; Tamaru, H.; Selker, E. U.; Cheng, X. *Mol. Cell* **2003**, *12*, 177–85.

(5) (a) Couture, J. F.; Collazo, E.; Brunzelle, J. S.; Trievel, R. C. *Genes Dev.* **2005**, *19* (12), 1455–1465. (b) Xiao, B.; Jing, C.; Kelly, G.; Walker, P. A.; Muskett, F. W.; Frenkiel, T. A.; Martin, S. R.; Sarma, K.; Reinberg, D.; Gamblin, S. J.; Wilson, J. R. *Genes Dev.* **2005**, *19*, 1444–1454.

(6) Xiao, B.; Jing, C.; Wilson, J. R.; Walker, P. A.; Vasisht, N.; Kelly, G.; Howell, S.; Taylor, I. A.; Blackburn, G. M.; Gamblin, S. J. *Nature* **2003**, *421*, 652–656.

(7) Couture, J. F.; Dirk, L. M. A.; Brunzelle, J. S.; Houtz, R. L.; Trievel, R. C. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 20659–20664.

(8) (a) Guo, H. B.; Guo, H. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 8797–802. (b) Xu, Q.; Chu, Y.-Z.; Guo, H.-B.; Smith, J. C.; Guo, H. *Chem.—Eur. J.* **2009**, *15*, 12596–12599.

(9) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.

(10) Field, M. J.; Bash, P. A.; Karplus, M. *J. Comput. Chem.* **1990**, *11*, 700–733.

(11) (a) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935. (b) Neria, E.; Fischer, S.; Karplus, M. *J. Chem. Phys.* **1996**, *105*, 1902–1921.

(12) Brooks, C. L.; Brunger, A.; Karplus, M. *Biopolymers* **1985**, *24*, 843–865.

(13) (a) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260–7268. (b) Cui, Q.; Elstner, M.; Kaxiras, E.; Frauenheim, T.; Karplus, M. *J. Phys. Chem. B* **2001**, *105*, 569–585.

(14) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(15) (a) Hu, P.; Wang, S.; Zhang, Y. *J. Am. Chem. Soc.* **2008**, *130*, 3806–3813. (b) Hu, P.; Zhang, Y. K. *J. Am. Chem. Soc.* **2006**, *128*, 1272–1278.

(16) Torrie, G. M.; Valleau, J. P. *Chem. Phys. Lett.* **1974**, *28*, 578–581.

(17) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011–1021.

(18) Zhang, X. D.; Bruice, T. C. *Biochem.* **2008**, *47*, 6671–6677.

(19) (a) Zhang, X.; Bruice, T. C. *Biochemistry* **2007**, *46*, 5505–14. (b) Zhang, X.; Bruice, T. C. *Biochemistry* **2007**, *46*, 14838–14844. (c) Zhang, X.; Bruice, T. C. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 5728–5732. (d) Zhang, X. D.; Bruice, T. C. *Biochemistry* **2007**, *46*, 9743–9751. (e) Zhang, X. D.; Bruice, T. C. *Biochemistry* **2008**, *47*, 2743–2748.

CT9006458

JCTC Journal of Chemical Theory and Computation

# Temperature Dependence of Protein Dynamics Simulated with Three Different Water Models

Dennis C. Glass,[†] Marimuthu Krishnan,[†] David R. Nutt,[‡] and Jeremy C. Smith*,[†]

*University of Tennessee/ORNL Center for Molecular Biophysics, Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, Tennessee 37831, and Department of Chemistry, University of Reading, Whiteknights, Reading RG6 6AD, United Kingdom*

Received December 3, 2009

**Abstract:** The effect of variation of the water model on the temperature dependence of protein and hydration water dynamics is examined by performing molecular dynamics simulations of myoglobin with the TIP3P, TIP4P, and TIP5P water models and the CHARMM protein force field at temperatures between 20 and 300 K. The atomic mean-square displacements, solvent reorientational relaxation times, pair angular correlations between surface water molecules, and time-averaged structures of the protein are all found to be similar, and the protein dynamical transition is described almost indistinguishably for the three water potentials. The results provide evidence that for some purposes changing the water model in protein simulations without a loss of accuracy may be possible.

## 1. Introduction

Water plays an important role in many chemical and biological processes.[1−9] For example, hydration water strongly influences the three-dimensional structure, dynamics, and function of proteins.[2] Water−protein interactions modify the free energy landscape that determines the folding, structure, and stability of proteins.[3−5] Internal protein dynamics, which are required for biological functions, are dependent on the level of hydration,[6] and dynamical coupling between the protein and water influences conformational flexibility.[7−10]

Protein hydration water can be grouped into two classes: internal water molecules, which can play structural and/or catalytic roles,[11,12] and external surface water molecules. Hydration water is experimentally estimated to account for 10−15% of the total cell water,[13,14] of which a small fraction of ∼0.1% is internal water molecules. Properties of water in the external hydration shell are modified compared to the bulk, with, for example, changes in average density[15−17] and perturbations in translational and rotational dynamics, and these changes have been extensively investigated using techniques such as neutron scattering,[13,18−25] nuclear magnetic resonance,[26−31] fluorescence spectroscopy,[32] mid-infrared pump−probe spectroscopy,[33] and molecular dynamics (MD) simulations.[21,33−40]

Empirical, molecular mechanics force fields, such as CHARMM,[41−43] AMBER,[44] GROMOS,[45] and OPLS-AA[46] are widely employed in atomistic MD simulations of biological molecules and, in order to represent the protein−solvent energy accurately, most molecular simulation applications employ explicit water models. A large number of water models is available. However, individual biomolecular force fields have normally been parametrized for use with a single water model (see e.g. ref 47) such that, during the parametrization, care can be taken to correctly balance water−water, water−protein, and protein−protein interactions. Nevertheless, the question arises as to whether the water−protein potentials are sufficiently robust so that alternative water models may be employed with any given biomolecular force field without a serious loss of accuracy. Flexibility in the choice of the water model may be of particular interest when the system is to be simulated at nonphysiological temperatures or pressures, as required, for example, in studies on antifreeze proteins,[48,49] under which circumstances different water models may exhibit significantly different properties (see e.g., ref 50), and/or when water properties are specifically under investigation for which an alternative water potential may be more accurate than the original. As another example, the dynamical equilibrium

* Corresponding author e-mail: smithjc@ornl.gov.

[†] Oak Ridge National Laboratory.

[‡] University of Reading.

10.1021/ct9006508 © 2010 American Chemical Society
Published on Web 03/04/2010

between water molecules on the protein surface and bulk is important in the study of the dielectric relaxation of aqueous protein solutions, and the rate of transition of water molecules from bulk to interface or vice versa is influenced by the choice of water model.[51] The need for alternate water models to better characterize the hydrophobic effect,[52] spectral properties,[53] solvation free energies,[54] and hydration dynamics[55,56] of proteins has also been discussed.

In recent work, the effects of varying the water model in molecular mechanics and dynamics calculations on the hydration of N-methylacetamide (NMA) and other small solute molecules and a small protein using the CHARMM force field[41−43] were examined with a focus on structural aspects, e.g., distribution functions around different biomolecular sites.[57] The overall description of solvation and biomolecular properties were found to be similar for the three models tested: TIP3P, TIP4P, and TIP5P.[50,58] The CHARMM protein force field was originally parametrized for use with the TIP3P potential. However, the results provide an indication that molecular simulations with the CHARMM force field may in some cases be performed with water models other than TIP3P.

In the present work, we compare the results of using the above three water models (TIP3P, 4P, 5P) on the temperature dependence of internal protein motion. Experimental and theoretical studies have found that proteins undergo a transition in internal dynamics at $T_g \approx 180-220$ K,[59−70] characterized by a rapid increase in the average protein atomic mean-square displacement above $T_g$. The transition at $T_g$ is strongly coupled to the solvent dynamics.[2,67−69,71−81] An additional low-temperature transition (∼150 K), that has been attributed to the activation of methyl group rotations, is present also in dehydrated proteins.[71,82−86] The dynamical transition at $T_g$ can be eliminated when dehydrating the protein or coating it with a bioprotectant,[87] and the value of $T_g$ can be shifted by changing the solvent composition.[88] MD simulations have demonstrated that the $T_g$ dynamical transition is driven by translational solvent dynamics.[68,89] For some proteins, the $T_g$ dynamical transition has also been correlated with protein activity,[64,66,73,90,91] although activity has also been observed for $T < T_g$.[92−95]

The question arises as to whether the dynamical properties of the protein are affected by the choice of the water model used. To investigate this, we have performed simulations of myoglobin at temperatures ranging from 20 to 300 K using the TIP3P, TIP4P, or TIP5P potentials and the CHARMM protein force field.[41−43] The protein dynamical transition was found to be unaffected by changes in the solvent model. Moreover, although the bulk properties of the three water models are markedly different, when interacting with the protein surface, the three water models behave similarly.

## 2. Methods

The set of models considered here is the TIP3P, TIP4P, and TIP5P family.[50,58] TIP3P is the standard model in the widely used CHARMM force fields. However, TIP4P and TIP5P are easy to implement for use with CHARMM and exhibit improved bulk water properties, as described below.

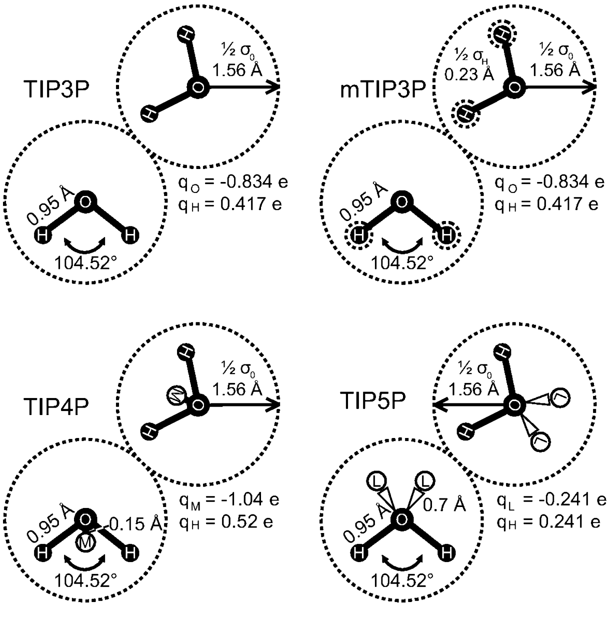


***Figure 1.*** TIPnP geometries. Dashed lines represent 1/2 of the van der Waals radius $\sigma_0$.

***Table 1.*** TIP3P, TIP4P, TIP5P, and mTIP3P Potential Energy Parameters

| | TIP3P | mTIP3P | TIP4P | TIP5P |
|---|---|---|---|---|
| $q_H$ | 0.417 | 0.417 | 0.520 | 0.241 |
| $q_O$ | −0.834 | −0.834 | | |
| $q_M$ | | | −1.040 | |
| $q_L$ | | | −0.241 | |
| $\sigma_{OO}$/A | 3.5364 | 3.5364 | 3.5399 | 3.5021 |
| $\varepsilon_O$/kcal/mol | 0.1521 | 0.1521 | 0.1550 | 0.1600 |
| $\sigma_{HH}$/A | | 0.4490 | | |
| $\varepsilon_H$/kcal/mol | | 0.0460 | | |
| $r_{OH}$/A | 0.9572 | 0.9572 | 0.9572 | 0.9572 |
| $r_{OM}$/A | | | 0.15 | |
| $r_{OL}$/A | | | | 0.7 |
| $\theta_{HOH}$/deg | 104.52 | 104.52 | 104.52 | 104.52 |
| $\theta_{LOL}$/deg | | | | 109.47 |

The TIP geometries and associated parameters are shown in Figure 1 and Table 1. In the TIP3P model, charges are placed at the hydrogen positions and on the oxygen, resulting in three interaction sites. A van der Waals term provides additional nonbonding interactions involving the oxygen atoms only.[58] Geometrical parameters were assigned according to experimental gas phase values.

mTIP3P is a slightly modified version of TIP3P, commonly used with CHARMM,[41−43] and includes additional van der Waals interaction sites at the hydrogen positions.[96] The effect of these additional terms on the properties of TIP3P has been shown to be small.[97] Therefore, in the following, TIP3P is used to refer to the CHARMM-modified version, mTIP3P.

The combination of TIP3P with the CHARMM force field has proven to be useful for examining the structure and dynamics of biomolecular systems. However, although TIP3P adequately describes the first hydration shell of bulk water, it lacks accuracy for the second hydration shell, for which the corresponding peak in the oxygen−oxygen radial distribution function is almost completely absent.[50] TIP4P, in which the oxygen charge site is moved along the HOH bisector toward the molecular center of mass, better reproduces experimental distribution functions than TIP3P.[58] In TIP5P, there are two lone-pair interaction sites, L, moved from the oxygen along the HOH bisector away from the hydrogens and symmetrically placed out of the HOH plane with $\angle_{\mathrm{LOL}} = 109.47°$.[50] The TIP4P and TIP5P oxygens carry no charge. TIP5P, explicitly incorporating tetrahedrality in the water model, is especially successful in reproducing bulk water density over a wide range of temperatures.[50]

Here, the CHARMM program package version c33b2 was used to perform MD simulations of hydrated myoglobin with the CHARMM22 force field and TIP3P, TIP4P, or TIP5P water.[41−43] The protein−water heteroatomic interaction parameters were calculated using the standard Lorentz−Berthelot mixing rule, which is defined as follows:

$$\sigma_{\mathrm{PW}} = \frac{\sigma_{\mathrm{P}} + \sigma_{\mathrm{W}}}{2}; \; \varepsilon_{\mathrm{PW}} = \sqrt{\varepsilon_{\mathrm{P}}\varepsilon_{\mathrm{W}}}$$

where ($\sigma_{\mathrm{P}}$, $\varepsilon_{\mathrm{P}}$) and ($\sigma_{\mathrm{W}}$, $\varepsilon_{\mathrm{W}}$) correspond to Lennard-Jones parameters associated with protein and water atoms, respectively. The 1.15 Å resolution myoglobin structure 1A6G, taken from the RCSB Protein Data Bank (www.pdb.org),[98] was used as the starting protein configuration. The model was constructed as in ref 68 to mimic a hydrated powder sample and thus model the experimental neutron scattering setup of ref 77, the data from which serve as a reference here. To do this, the protein was solvated by placing it in a box of water, retaining only those 492 molecules closest to the protein. The temperature dependent dynamical properties in the present simulations were found to be very similar to those derived from NPT simulations in solution on the same myoglobin structure.[86]

Electrostatic interactions were truncated using a shift function with a 12 Å cutoff, and a switch function was used for the truncation of van der Waals interactions between 10 and 12 Å. The SHAKE algorithm was used to constrain all bond lengths involving hydrogens.[99] The structures were energy minimized using 600 steepest descent steps and 2500 conjugate gradient steps with the protein atoms fixed, then with fixed solvent allowing the protein to move, and finally without any constraints.

After heating to the desired temperature in steps of 5 K every 1000 dynamics steps, the system was equilibrated for 400 ps. Subsequently, 1 ns production runs were performed in the NVT ensemble at temperatures from 20 to 300 K in intervals of 20 K and in smaller intervals of 10 K between 120 and 240 K. The system was kept at constant temperature using the weak coupling algorithm of ref 100. A time step of 1 fs was used for the integration of the equations of motion. Coordinates and velocities were saved every 50 steps. The simulation protocol was the same for all temperatures and water models.

In order to examine in more detail the influence of the water model on the time averaged protein structure at 300 K, 10 additional simulations of 1 ns length were performed for each water model, starting with different, randomly chosen, initial velocity assignments.

To avoid potential artifacts, no restraining potential was applied to the water molecules to prevent evaporation. Subsequently, for $T > 260$ K, a small number of molecules evaporated from the water shell surrounding the protein, and these were excluded from all analyses. No evaporation occurred for $T < 260$ K, including during extended 7 ns simulations with TIP3P at 180, 220, and 260 K.

In addition to the above calculations, for comparison purposes, simulations of bulk water using the TIP water models were performed. For these, GROMACS 4.0[101] was used to generate a 30 Å cubic water box at 300 K and 1 atm pressure containing 895 molecules for TIP3P, 886 for TIP4P, and 878 for TIP5P. The electrostatic interactions were treated in the same way as for the above protein simulations, and periodic boundary conditions were applied. The configurations were energy minimized using 600 steepest descent steps, equilibrated for 500 ps at the desired temperatures and, subsequently, 1 ns production runs were performed in the NVT ensemble at temperatures from 20 to 300 K in intervals of 20 K and in smaller intervals of 10 K between 100 and 260 K. The system was kept at a constant temperature using the weak coupling algorithm of ref 100. A time step of 2 fs was used. Coordinates were saved every 50 steps. The simulation protocol was again the same for all temperatures and water models.

Additional MD simulations of myoglobin fully solvated in a periodic TIP3P water box ($68 \times 68 \times 52$ Å$^3$ dimensions) were carried out using the NAMD software.[102] The Particle Mesh Ewald (PME) method was used for the electrostatics, and a switch function was used for truncation of van der Waals interactions between 10 and 12 Å. The starting configurations were energy minimized using 6000 conjugate gradient steps followed by 1 ns equilibration and 1 ns production runs in the NPT ensemble. The equations of motion were integrated with a time step of 1 fs, and the atomic coordinates were saved at every 100 fs. Using the Langevin thermostat and barostat, simulations were carried out at a range of temperatures between 20 and 300 K and at 1 atm of pressure. The temperature dependence of the Kirkwood g-factor determined from this set of simulations serves for comparison with results obtained from the hydrated powder model.

## 3. Results

**3.1. Time-Averaged Structures.** The influence of the water models on the average structure of myoglobin is first investigated. The time-averaged structures were calculated for all 30 independent simulations at 300 K and were compared using a root-mean-square deviation (RMSD) per residue, defined here as

$$\mathrm{RMSD}_i^{A,B} = \sqrt{\frac{\sum_{j=1}^{N_i} (x_j^A - x_j^B)^2}{N_i}} \tag{1}$$

where $i$ is the residue number, $N_i$ is the number of protein heavy atoms in residue $i$, $A$ and $B$ are any two given time-averaged structures, and $x_j^{A/B}$ denotes the heavy atom coordinates. Convergence was checked by comparing RMSD values from the first and second half of each simulation, which were found to be in close agreement.

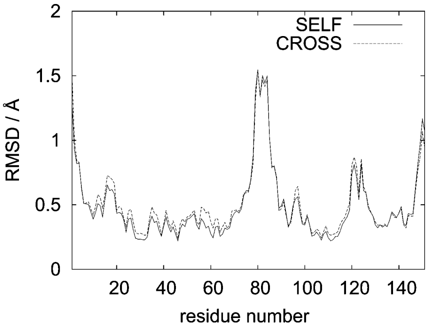


***Figure 2.*** Backbone heavy-atom RMSD per residue at 300 K. Average over pairs of structures solvated in the same or in different water models. “SELF” refers to RMSD values between protein structures solvated by the same water model, whereas “CROSS” refers to deviations in configurations between the models.

Figure 2 shows the backbone heavy-atom RMSD per residue based on the average over all pairs ($A$,$B$) of structures solvated in the same or in different water models. “SELF” refers to RMSD values between protein structures solvated by the same water model, whereas “CROSS” refers to deviations in configurations between the models, for example, structures of the protein in TIP3P compared with structures in TIP4P or TIP5P.

Neglecting the five C- and N-terminal residues, the $\mathrm{RMSD}_i^{A,B}$ averaged over residues is 0.47 ± 0.25 Å, averaged over “SELF” is 0.46 ± 0.26 Å, and averaged over “CROSS” is 0.49 ± 0.25 Å, and the average difference between “SELF” and “CROSS” is 0.035 ± 0.026 Å. Therefore, the RMSD per residue is similar for both the “SELF” and “CROSS” data sets, indicating that variation of the water model does not significantly influence the time-averaged protein RMSD.

**3.2. Mean-Square Displacement.** The mean-square displacement $\langle r^2(t)\rangle$ is defined as

$$\langle r^2(t)\rangle = \left\langle \frac{1}{N}\sum_{i=0}^{N}(r_i(t+t_0) - r_i(t_0))^2 \right\rangle_{i,t_0} \qquad (2)$$

where $N$ is the number of atoms, $(r_i(t + t_0) - r_i(t_0))$ is the displacement of atom $i$ in time $t$, and $\langle \bullet \rangle_{i,\,t_0}$ represents the ensemble average, approximated as a time average over $t_0$ by assuming ergodicity.

The mean-square displacement (MSD) averaged over the heavy atoms of myoglobin was calculated for all temperatures and water models. Figure 3C shows the time evolution of heavy atom MSD of myoglobin at various temperatures, while Figure 3A exhibits time-averaged MSD as a function of temperaure. The inset in Figure 3 A shows an expanded view of the low-temperature region, up to $T \approx 210$ K. The MSD rises linearly for low temperatures until a first change in the gradient at $T \approx 150$ K, which has been attributed to the activation of methyl group rotations.[71,84,86] At $T \approx 220$ K, $\langle r^2(t)\rangle$ exhibits a further increase in gradient corresponding to the solvent-driven dynamical transition. $\langle r^2(t)\rangle$, both for <220 K and >220 K, is similar using all three TIP water models.

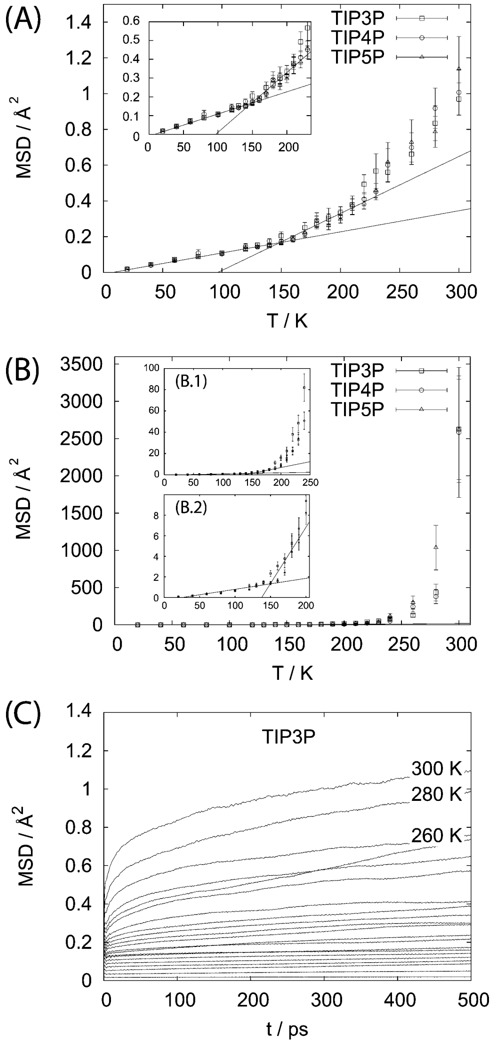


***Figure 3.*** (A) Heavy atom mean-square displacement of myoglobin solvated in TIP3P, TIP4P, or TIP5P as a function of temperature. The inset shows an expanded view of the low temperature region up to $T \approx 210$ K. (B) Mean-square displacement of the hydration shell water as a function of temperature. The insets (B.1 and B.2) give expanded views since the data cover multiple orders of magnitude. (C) Time dependent mean-square displacement of myoglobin in TIP3P as a function of time at different temperatures.

Figure 3B shows the $\langle r^2(t)\rangle$ of the water molecules in the protein simulations as a function of temperature. The insets B.1 and B.2 give expanded views since the data cover multiple orders of magnitude. The profiles qualitatively resemble Figure 3A; i.e., the MSD for the water molecules rises linearly for low temperatures until, interestingly, a first change in slope is seen at $T \approx 150$ K (Figure 3B.2) followed by a second transition at $T \approx 200$ K (Figure 3B.1).

The transition at 220 K in the protein has been previously observed and the strong coupling between the solvent and protein characterized.[62,65,68,73,76,78,80,89,103−109] However, the low-$T$ transition at 150 K for water was unexpected, and the question therefore arises as to whether it would occur independent of the protein. To check this, a set of simulations of pure TIP water was performed, as described in the Methods. Figure 4A shows the TIP water mean-square displacements as a function of temperature. A comparison between Figures 3B and 4A indicates that, while the dynamical transition behavior of the protein and protein

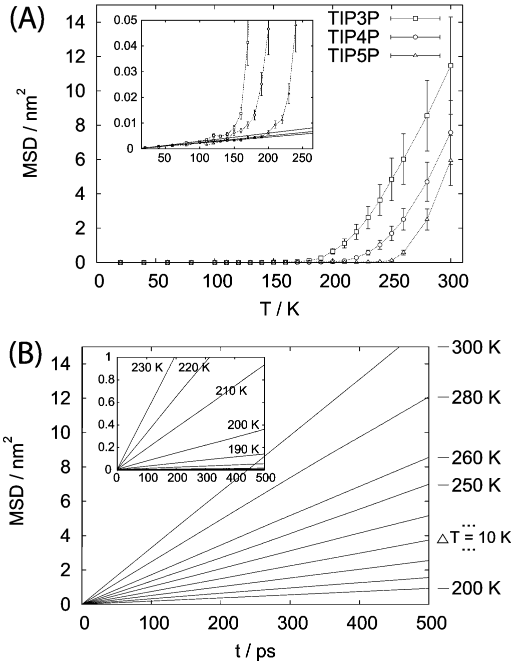


***Figure 4.*** (A) Mean-square displacement of TIP3P, TIP4P, or TIP5P water molecules as a function of temperature from bulk water simulations. The inset shows an expanded view of the low temperature region. The solid lines corresponds to linear fits for $T = 20-100$ K. (B) Mean-square displacement as a function of time for TIP3P and various temperatures. The insets show data belonging to further temperatures in an appropriate magnification of the abscissa.

hydration water are similar, they do not resemble that of the bulk water models.

The data for the three TIP models exhibit similar properties in that $\langle r^2(t)\rangle$ is linear for $T \leq T_1$. However unlike in the solution simulations, the temperature at which the slope changes, $T_1$, is markedly different for the three models with $T_1 \approx 120$ K for TIP3P, 140 K for TIP4P, and 190 K for TIP5P. Further, again unlike the data for the hydrated protein, the bulk data do not show two distinct changes in slope.

The self-diffusion constant, $D$, was calculated from the linear part of each mean-squared displacement, at, i.e., $t >$ 200 ps (see Figure 4B), using the Einstein relation:

$$\lim_{t\to\infty}\langle |r(t_0 + t) - r(t_0)|^2\rangle = 6Dt \tag{3}$$

where $r(t)$ is the position of the water oxygen atom and $\langle\cdot\rangle$ denotes averaging over both time origins $t_0$ and the water molecules. $D$, in units of $10^{-5}$ cm$^2$/s, at 300 K is found to be 5.46 ± 0.02 (TIP3P), 3.60 ± 0.06 (TIP4P), and 2.82 ± 0.02 (TIP5P). Corresponding values in the literature vary depending on the simulation setup, but reported values using a similar protocol to that used here (i.e., shift electrostatics) in the NPT ensemble are 5.8 ± 0.2 (TIP3P), 3.78 ± 0.02 (TIP4P), and 2.94 ± 0.06 (TIP5P),[110] in good agreement with the present work. For completeness, in Table 2, $D$ is also given for the other temperatures.

**3.3. Heterogeneous Distribution of Anharmonic Motions among Protein Residues.** To quantify the fraction of protein residues that exhibit large anharmonic dynamics, the MSD per residue ($\langle r^2(t)\rangle_i$, where $i$ denotes the residue index) was calculated. $\langle r^2(t)\rangle_i$ is decomposed into harmonic and anharmonic components as follows:

$$\langle r^2(t)\rangle_i = \langle r^2(t)\rangle_{i,\text{harmonic}} + \langle r^2(t)\rangle_{i,\text{anharmonic}} \tag{4}$$

where $\langle r^2(t)\rangle_{i,\text{harmonic}}$ is obtained by linearly fitting $\langle r^2(t)\rangle_i$ for $T \leq 140$ K and extrapolating to higher $T$.

A residue is denoted as exhibiting anharmonic dynamics if $\langle r^2(t)\rangle_i > (\langle r^2(t)\rangle_{i,\text{harmonic}} + 2\sigma)$, where $\sigma$ is the standard deviation at the onset of anharmonicity (140 K). As a control, it was determined whether the onset of anharmonicity depends on the temperature interval chosen for the estimation of harmonic dynamics: using 20−100 K as the fitting interval and $\sigma_{100\text{K}}$, the procedure was found to give similar results.

Figure 5 shows the temperature dependence of the fraction of residues exhibiting anharmonic dynamics. For $T \leq 140$ K, almost all residues exhibit only harmonic motion, with $\langle r^2(t)\rangle_i$ similar to each other, while an abrupt change is evident at higher $T$, as an increasing fraction of residues exhibits anharmonic dynamics. Even well above the dynamical transition temperature at 260 K, approximately 25% of the residues still remain harmonic. This result is consistent with previous simulation work on myoglobin, in which the onset of anharmonicity was found to be gradual with $T$.[69] Again, when comparing the protein simulations using different water models, there are no statistically significant differences.

**3.4. Reorientational Relaxation Time.** The rotational dynamics of water can be characterized by the water dipole orientation autocorrelation function, $C(t)$:

$$C(t) = \langle \vec{e}_i(t + t_0)\cdot \vec{e}_i(t_0)\rangle_{i,t_0} \tag{5}$$

where $\vec{e}_i(t)$ is a unit vector along the water dipole. For liquids, $C(t)$ decays to zero as the dipole loses its memory of its initial orientation.

Since the structural and dynamic properties of the hydration layer water molecules depend on the heterogeneous surface roughness and charge distribution,[21,24,39,40,111] a multiexponential decay is expected. From the protein simulations, the reorientational relaxation time of water, $\tau$, was calculated by fitting the following triple exponential function to $C(t)$:

$$\begin{aligned} C(t) = \; & a_0 \exp(-t/\tau_0) + a_1 \exp(-t/\tau_1) + \\ & (1 - a_0 - a_1)\exp(-t/\tau_2) \end{aligned} \tag{6}$$

Equation 6 was found to capture the decay in the target data, whereas the fitting procedure failed for simpler fitting functions. The relaxation time, $\tau$, was derived using the following relation:

$$\tau = a_0\tau_0 + a_1\tau_1 + (1 - a_0 - a_1)\tau_2 \tag{7}$$

Figure 6A shows $C(t)$ together with the fit of eq 7 for TIP3P hydration water at different temperatures. $C(t)$'s for TIP4P and TIP5P hydration water exhibit similar decay behaviors (not shown). In general, the profiles consist of a fast decay on the picosecond time scale followed by slower dynamics. For most temperatures, $C(t)$ does not fully decay to zero on the present time scale of ~500 ps. For temperatures $T > 220$

***Table 2.*** Diffusion Constants in Units of $10^{-5}$ cm$^2$/s Calculated Using the Einstein Relation for TIP3P, TIP4P, and TIP5P from a 30 Å Cubic Water Box with the Same Electrostatic Treatment As in the Protein Simulation

| *T*/K | TIP3P | | | TIP4P | | | TIP5P | | |
|---|---|---|---|---|---|---|---|---|---|
| 150 | 0.0018 | ± | 0.0000 | 0.0018 | ± | 0.0002 | 0.0007 | ± | 0.0001 |
| 160 | 0.0043 | ± | 0.0002 | 0.0016 | ± | 0.0002 | 0.0007 | ± | 0.0002 |
| 170 | 0.0171 | ± | 0.0007 | 0.0025 | ± | 0.0002 | 0.0009 | ± | 0.0001 |
| 180 | 0.0444 | ± | 0.0033 | 0.0036 | ± | 0.0001 | 0.0009 | ± | 0.0000 |
| 190 | 0.116 | ± | 0.0045 | 0.0086 | ± | 0.0011 | 0.0009 | ± | 0.0001 |
| 200 | 0.3104 | ± | 0.0075 | 0.0193 | ± | 0.0004 | 0.0014 | ± | 0.0000 |
| 210 | 0.5058 | ± | 0.0146 | 0.04 | ± | 0.0007 | 0.0028 | ± | 0.0004 |
| 220 | 0.8463 | ± | 0.0032 | 0.1306 | ± | 0.0026 | 0.003 | ± | 0.0001 |
| 230 | 1.243 | ± | 0.0177 | 0.2699 | ± | 0.0010 | 0.0073 | ± | 0.0016 |
| 240 | 1.7191 | ± | 0.0464 | 0.4948 | ± | 0.0044 | 0.0202 | ± | 0.0002 |
| 250 | 2.3816 | ± | 0.0269 | 0.8021 | ± | 0.0017 | 0.0459 | ± | 0.0072 |
| 260 | 2.8328 | ± | 0.0261 | 1.1895 | ± | 0.0040 | 0.2601 | ± | 0.0001 |
| 280 | 3.9614 | ± | 0.1640 | 2.1882 | ± | 0.0674 | 1.1849 | ± | 0.0100 |
| 300 | 5.4596 | ± | 0.0158 | 3.595 | ± | 0.0581 | 2.8162 | ± | 0.0190 |

K (above the dynamical transition temperature), $C(t)$ decays rapidly (<50 ps) to ~0.5, but the decay is much slower at lower $T$.

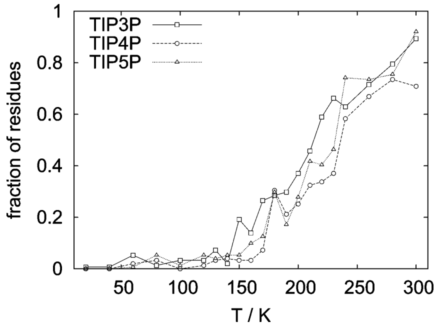


***Figure 5.*** Fraction of residues exhibiting anharmonic dynamics.

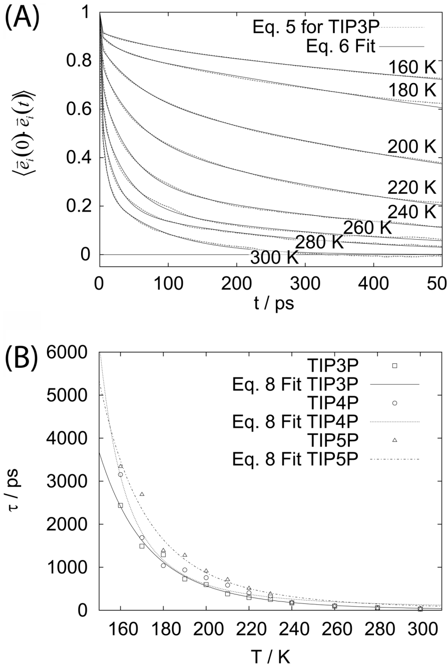


***Figure 6.*** (A) Protein hydration water dipole orientational autocorrelation functions (eq 5) for TIP3P and eq 6 fitted to data from eq 5. TIP4P and TIP5P data are similar. (B) Reorientational relaxation lifetimes $\tau$ (eq 7) fitted with the Vogel−Fulcher−Tammann equation (eq 8).

There is good agreement between the fitted curves of eq 6 and the data (Figure 6A). Figure 7 and Table 3 present the temperature dependence of the resulting fitted parameters. At low $T$ ($\leq$ 180 K), the process associated with relaxation time $\tau_2$ dominates, with a weight factor, $1 - a_0 - a_1 \approx 1$. A decrease of $1 - a_0 - a_1$ (i.e., an increase of $a_0$ and/or $a_1$) is observed for $T > 180$ K, indicating the activation of additional relaxation processes. For $T \geq 240$ K, the three components are approximately equally weighted.

The temperature dependences of the weights and values of $\tau_0$ and $\tau_1$ behave very similarly to each other, with the weight of the $\tau_1$ component being larger and the values being ~10 times larger than $\tau_0$. $\tau_0$ and $\tau_1$ have broad maxima at ~260 and 210 K, respectively. For temperatures $T \leq 180$ K, the $\tau_1$ component jumps to high values, accompanied by a drop in its weight.

The relaxation times $\tau$, derived by eq 7, are given in Figure 6B. Relaxation times at 300 K from the full sets of simulations at this temperature are 31 ± 3, 35 ± 3, and 32 ± 6 ps for TIP3P, TIP4P, and TIP5P, respectively, and rise by 2 orders of magnitude as the temperature decreases to 160 K.

At temperatures $T > T^*$, glass-forming liquids exhibit an Arrhenius relaxation mechanism $\tau \propto \exp(-E_\beta/k_BT)$ due to the behavior of the $\beta$ relaxation which, at $T^*$, splits into a fast $\beta$ relaxation and a slow α relaxation.[112,113] The α relaxation may often be described with a Vogel−Fulcher−Tammann (VFT) equation over the range $T_g < T < T^*$, i.e.,

$$\tau(T) = \tau_c \exp\left(\frac{A}{T - T_0}\right) \tag{8}$$

with fitting parameters $\tau_c$, $A$, and $T_0$. Cooperativity of $\beta$-relaxation events has been suggested as the origin of the α relaxation. A similar VFT relationship between $\tau$ and $T$ has been used in glass physics,[112] where $T_0$ has been hypothesized to be the Kauzmann temperature and $A = (E_\beta/R)[(T^* - T_0)/T^*]$ with $E_\beta$ as the activation energy for the $\beta$ relaxations. The Kauzmann temperature is the temperature at which the entropies of the supercooled liquid and the corresponding crystal are in principle equal.[113,114]

Angell proposed a "fragile-to-strong" classification of liquids in which relaxation times of "strong" liquids follow an Arrhenius trend (e.g., $SiO_2$), whereas “fragile” liquids deviate from such a behavior.[115] The VFT relation describes well the present data for $T > 160$ K, although the fitted parameter values were subject to large errors, thus classifying the protein hydration shell water as a “fragile” liquid in Angell’s scheme.

**3.5. Local Orientational Ordering.** The local orientational ordering of water dipoles can be quantified by the distance-dependent Kirkwood $G$ factor, defined as follows:[116]

$$G_{K,h}(r) = \langle \vec{\mu}_i \cdot \vec{M}(r) \rangle \qquad (9)$$

with

$$\vec{M}(r) = \sum_{r_{ij} \leq r} \vec{\mu}_j \qquad (10)$$

where $\vec{\mu}_i$ denotes a unit vector in the direction of the dipole moment of molecule $i$. $G_{K,h}(r)$ is equal to two if a pair of water dipoles is parallel. Elevated $G_{K,h}(r)$ therefore corresponds to high angular correlations between water molecules.

Figure 8 shows $G_{K,h}^{r_{max}\pm0.2}(T)$, as a function of temperature for different water models, averaged over 0.2 Å around $r_{max}$, where $r_{\max}$ is the most probable near neighbor distance, taken to be equal to the position of the first peak in the water oxygen−oxygen radial distribution function $g_{OO}(r)$ (~2.8 Å). Other simulation studies obtained similar values for $G_{K,h}^{r_{\max}}(T)$ at ~2.8 Å and 300 K.[117,118]

A systematic decrease in $G_{K,h}^{r_{\max}\pm0.2}(T)$ is evident above 200 K for all water models in the hydration layer of myoglobin (in the hydrated power model), associated with the increased diffusion (see Figure 3B). Although $G_{K,h}^{r_{\max}\pm0.2}(T)$ obtained for a myoglobin solution is slightly higher than that of the hydrated powder model, the temperature dependence of $G_{K,h}^{r_{\max}\pm0.2}(T)$ shows a similar trend to that of the hydrated powder model. The observed higher values of $G_{K,h}^{r_{\max}\pm0.2}(T)$ in myoglobin solution can be attributed to the fact that all the water molecules (both bulk and interfacial) have complete first coordination shells in solution, while in the powder model some can have partial coordination shells. The results reported in Figures 3A,B and 8 are consistent with dynamical coupling between the protein and the solvent since the transition consistently occurs at ~220 K, captured by all TIP water models investigated here.

## 4. Conclusions

The effect of the variation of the water model on the temperature dependence of protein and hydration water dynamics has been investigated here by performing molecular dynamics simulations of hydrated myoglobin. Both protein and water properties were analyzed, including the time-averaged structures of the protein, the average mean-square displacements of the protein and water atoms, the solvent reorientational relaxation times, and pair angular correlations between the water molecules.

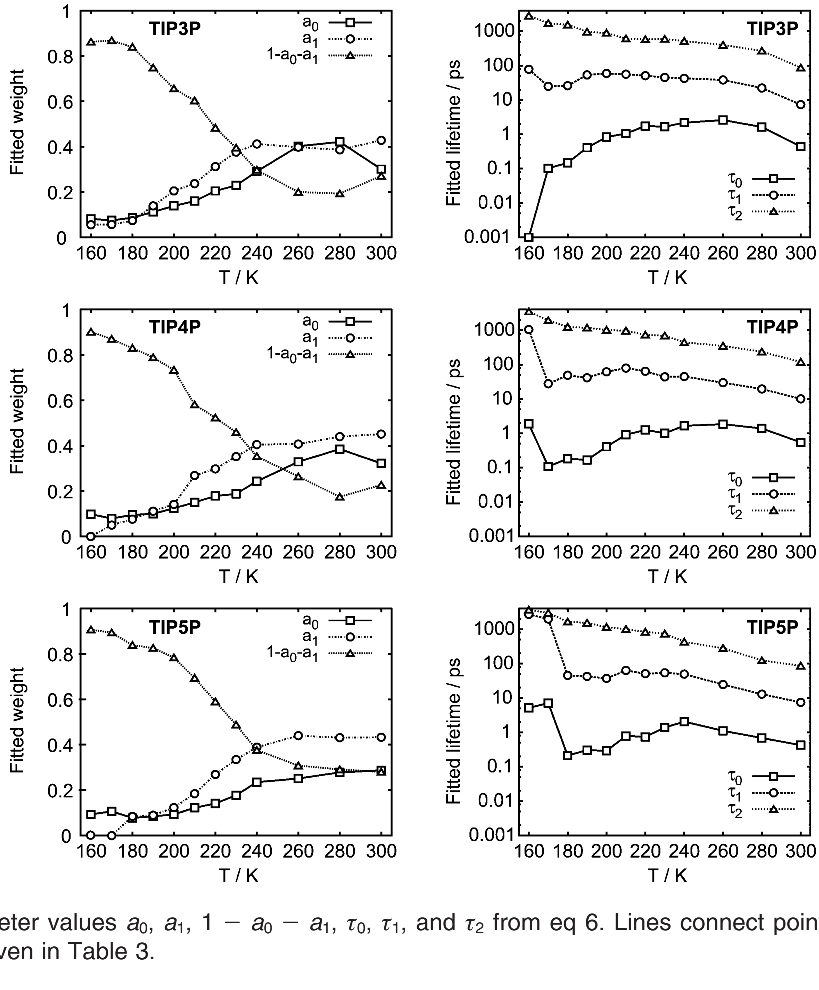


***Figure 7.*** Fitted parameter values $a_0$, $a_1$, $1 - a_0 - a_1$, $\tau_0$, $\tau_1$, and $\tau_2$ from eq 6. Lines connect points as a guide to the eye. Numerical values are given in Table 3.

***Table 3.*** Fitted Parameter Values for eq 6[a]

| | $T$/K | TIP3P | TIP4P | TIP5P | | $T$/K | TIP3P | TIP4P | TIP5P |
|---|---|---|---|---|---|---|---|---|---|
| $a_0$ | 160 | 0.08 | 0.10 | 0.09 | $\tau_0$ | 160 | 0.00 | 1.88 | 5.11 |
| | 170 | 0.08 | 0.08 | 0.11 | | 170 | 0.10 | 0.11 | 7.15 |
| | 180 | 0.09 | 0.09 | 0.08 | | 180 | 0.15 | 0.18 | 0.21 |
| | 190 | 0.11 | 0.10 | 0.08 | | 190 | 0.41 | 0.17 | 0.31 |
| | 200 | 0.14 | 0.12 | 0.09 | | 200 | 0.83 | 0.41 | 0.29 |
| | 210 | 0.16 | 0.15 | 0.12 | | 210 | 1.06 | 0.91 | 0.78 |
| | 220 | 0.21 | 0.18 | 0.14 | | 220 | 1.74 | 1.26 | 0.74 |
| | 230 | 0.23 | 0.19 | 0.18 | | 230 | 1.66 | 1.01 | 1.38 |
| | 240 | 0.29 | 0.24 | 0.24 | | 240 | 2.18 | 1.66 | 2.06 |
| | 260 | 0.40 | 0.33 | 0.25 | | 260 | 2.60 | 1.85 | 1.11 |
| | 280 | 0.42 | 0.38 | 0.28 | | 280 | 1.63 | 1.39 | 0.69 |
| | 300 | 0.30 | 0.32 | 0.29 | | 300 | 0.44 | 0.55 | 0.43 |
| $a_1$ | 160 | 0.06 | 0.00 | 0.00 | $\tau_1$ | 160 | 78.22 | 1049.77 | 2710.67 |
| | 170 | 0.06 | 0.05 | 0.00 | | 170 | 25.05 | 27.77 | 1984.47 |
| | 180 | 0.07 | 0.08 | 0.08 | | 180 | 26.10 | 48.92 | 45.39 |
| | 190 | 0.14 | 0.11 | 0.09 | | 190 | 53.70 | 41.52 | 42.55 |
| | 200 | 0.20 | 0.14 | 0.12 | | 200 | 59.20 | 61.70 | 37.20 |
| | 210 | 0.24 | 0.27 | 0.18 | | 210 | 56.04 | 79.61 | 63.44 |
| | 220 | 0.31 | 0.30 | 0.27 | | 220 | 50.97 | 64.38 | 51.01 |
| | 230 | 0.38 | 0.35 | 0.33 | | 230 | 45.34 | 44.00 | 53.98 |
| | 240 | 0.41 | 0.40 | 0.39 | | 240 | 42.57 | 44.90 | 49.49 |
| | 260 | 0.40 | 0.41 | 0.44 | | 260 | 38.17 | 29.87 | 24.68 |
| | 280 | 0.39 | 0.44 | 0.43 | | 280 | 22.27 | 19.51 | 12.89 |
| | 300 | 0.43 | 0.45 | 0.43 | | 300 | 7.32 | 10.09 | 7.46 |
| $1 - a_0 - a_1$ | 160 | 0.86 | 0.90 | 0.91 | $\tau_2$ | 160 | 2819.49 | 3498.03 | 3679.43 |
| | 170 | 0.87 | 0.87 | 0.89 | | 170 | 1714.96 | 1939.45 | 3007.36 |
| | 180 | 0.84 | 0.83 | 0.84 | | 180 | 1535.37 | 1245.80 | 1650.42 |
| | 190 | 0.75 | 0.79 | 0.83 | | 190 | 963.95 | 1182.51 | 1538.97 |
| | 200 | 0.66 | 0.73 | 0.78 | | 200 | 890.55 | 1020.11 | 1156.62 |
| | 210 | 0.60 | 0.58 | 0.69 | | 210 | 610.75 | 968.10 | 1007.51 |
| | 220 | 0.48 | 0.52 | 0.59 | | 220 | 580.03 | 736.41 | 838.47 |
| | 230 | 0.40 | 0.46 | 0.49 | | 230 | 594.33 | 690.84 | 741.91 |
| | 240 | 0.30 | 0.35 | 0.38 | | 240 | 515.06 | 443.10 | 431.47 |
| | 260 | 0.20 | 0.26 | 0.31 | | 260 | 398.39 | 349.69 | 280.67 |
| | 280 | 0.19 | 0.18 | 0.29 | | 280 | 269.09 | 235.99 | 123.27 |
| | 300 | 0.27 | 0.23 | 0.28 | | 300 | 86.83 | 119.99 | 86.76 |

[a] $\tau_1$, $\tau_2$, and $\tau_3$ in ps.

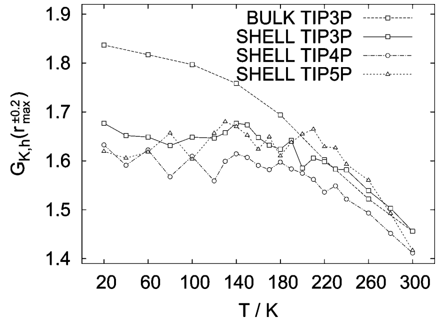


***Figure 8.*** Local orientational ordering of water dipoles measured with the distance-dependent Kirkwood *G*-factor $G_{K,h}^{r_{max}\pm0.2}(T)$ averaged over 0.2 Å around the maximum of the oxygen−oxygen radial distribution function as a function of temperature for different water models in the hydration shell ("SHELL") and for "BULK" TIP3P. $G_{K,h}^{r_{max}\pm0.2}(T)$ for "BULK" TIP3P is scaled by $c = 0.89$ (see text). Lines connect points as a guide to the eye.

Variation of the water model between TIP3P, TIP4P, and TIP5P leads to the same time-averaged structures of the protein to within statistical error. Also, for all three water models, the temperature-dependent mean-square displacement exhibits the well-known dynamical transition at $T_g \approx 220$ K. Furthermore, it has been previously shown that the SPC/E water model[119] with the GROMOS protein force field[120] also reproduces the $T_g$ protein dynamical transition at the same temperature.[103]

Mid-infrared pump−probe spectroscopy on the dynamics of HDO in $H_2O$ yielded an orientational lifetime of $\tau_{PP} = 2.5$ ps for bulk water and $\tau_{PP} > 10$ ps for "immobilized" water in the solvation shell of tetramethylurea,[33,121] which corresponds to a retardation factor of at least 4. The bulk water relaxation time $\tau_D$ was derived to be 8.8 ps from molecular dynamics studies and 7 ps from experiments on dielectric relaxation, both at 303 K.[122] [Pump−probe experiments measure different lifetimes than dielectric relaxation experiments or the present simulation data: whereas the lifetime $\tau_{PP}$ from pump−probe experiments is related to the second-order correlation function $\langle P_2(\cos\,\theta(t))\rangle$ (where $P_2$ is the second-order Legendre polynomial), the lifetimes $\tau_D$ from dielectric relaxation and the present work are determined by the first-order correlation function $\langle P_1(\cos\,\theta(t))\rangle$. The ratio between $\tau_{PP}$ and $\tau_D$ depends on the details of molecular diffusion: $\tau_{PP} = 3\tau_D$ for simple rotational diffusion but is different for other types of dynamics.[123]]

Experiments and simulations have shown that the reorientational dynamics of protein hydration water is slowed down relative to the bulk, e.g.,[5,26,27,34,36,40] with the slowing down being influenced by heterogeneous surface roughness and charge distribution.[21,24,39,40,111]

A recent measurement of cell water dynamics estimates the rotational correlation time for water directly interacting with biomolecular surfaces to be 27 ps.[14] The present protein hydration water relaxation times at 300 K are 31 ± 3, 35 ± 3, and 32 ± 6 ps for TIP3P, TIP4P, and TIP5P, respectively. These times are very similar to each other and to the values in ref 14 and are approximately a factor of 4 slower then the bulk relaxation times measured in ref 122.

A decrease in the pair angular correlations between water molecules in the protein hydration layer, calculated using the Kirkwood *G* factor, accompanies the dynamical transition.

Taken together, the present results suggest that the global dynamical properties of the protein and hydration water are not significantly affected by variation of the water models among TIP3P, TIP4P, and TIP5P. Although these models have documented different bulk phase properties, they behave similarly on the protein surface for the quantities investigated.

Broadly speaking, the present work has served two purposes. First, the relative invariance of the simulation-derived temperature-dependent protein and water dynamics to the water potential used bolsters the argument for the reliability of the simulation analyses of these phenomena previously published using only one water potential.[8,9,41−43,48,57,68−70,77,86,96,124−127] Second, the results indicate that, although general interchangeability of water potentials in protein simulations is not expected and cannot be assumed, for some purposes at least it is safe to choose between TIP3P, TIP4P, and TIP5P as the water potential used in protein simulations with CHARMM.

**Acknowledgment.** This work was funded by a U.S. Department of Energy Laboratory—Directed Research and Development grant.

## References

(1) Finney, J. L. *J. Mol. Liq.* **2001**, *90*, 303–312.

(2) Chaplin, M. *Nat. Rev. Mol. Cell Biol.* **2006**, *7*, 861–866.

(3) Kauzmann, W. *Adv. Protein Chem.* **1959**, *14*, 1–63.

(4) Dill, K. *Biochemistry* **1990**, *29*, 7133–7155.

(5) Halle, B. *Philos. Trans. R. Soc. London, Ser. B* **2004**, *359*, 1207–1224.

(6) Paciaroni, A.; Cinelli, S.; Cornicchi, E.; Francesco, A.; Onori, G. *Chem. Phys. Lett.* **2005**, *410*, 400–403.

(7) Nakasako, M. *Philos. Trans. R. Soc. London, Ser. B* **2004**, *359*, 1191–1206.

(8) Oleinikova, A.; Smolin, N.; Brovchenko, I.; Geiger, A.; Winter, R. *J. Phys. Chem. B* **2005**, *109*, 1988–1998.

(9) Smolin, N.; Oleinikova, A.; Brovchenko, I.; Geiger, A.; Winter, R. *J. Phys. Chem. B* **2005**, *109*, 10995–11005.

(10) Zanotti, J. M.; Bellissent-Funel, M. C.; Parello, J. *Biophys. J.* **1999**, *76*, 2390–2411.

(11) Williams, M. A.; Goodfellow, J. M.; Thornton, J. M. *Protein Sci.* **1994**, *3*, 1224–1235.

(12) Denisov, V. P.; Halle, B. *Faraday Discuss.* **1996**, *103*, 227–244.

(13) Stadler, A. M.; Embs, J. P.; Digel, I.; Artmann, G. M.; Unruh, T. B.

(14) Persson, E.; Halle, B. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 6266–6271.

(15) Svergun, D. I.; Richard, S.; Koch, M. H. J.; Sayers, Z.; Kuprin, S.; Zaccai, G. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 2267–2272.

(16) Merzel, F.; Smith, J. C. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 5378–5383.

(17) Merzel, F.; Smith, J. C. *J. Chem. Inf. Model.* **2005**, *45*, 1593–1599.

(18) Zanotti, J. M.; Bellissent-Funel, M. C.; Chen, S. H. *Phys. Rev. E* **1999**, *59*, 3084–3093.

(19) Bellissent-Funel, M. C.; Chen, S. H.; Zanotti, J. M. *Phys. Rev. E* **1995**, *51*, 4558–4569.

(20) Dellerue, S.; Bellissent-Funel, M. C. *Chem. Phys.* **2000**, *258*, 315–325.

(21) Russo, D.; Hura, G.; Head-Gordon, T. *Biophys. J.* **2004**, *86*, 1852–1862.

(22) Takahara, S.; Sumiyama, N.; Kittaka, S.; Yamaguchi, T.; Bellissent-Funel, M. *J. Phys. Chem. B* **2005**, *109*, 11231–11239.

(23) Jansson, H.; Bergman, R.; Swenson, J. *J. Non-Cryst. Solids* **2006**, *352*, 4410–4416.

(24) Malardier-Jugroot, C.; Johnson, M. E.; Murarka, R. K.; Head-Gordon, T. *Phys. Chem. Chem. Phys.* **2008**, *10*, 4903–4908.

(25) Frölich, A.; Gabel, F.; Jasnin, M.; Lehnert, U.; Österhelt, D.; Stadler, A. M.; Tehei, M.; Weik, M.; Wood, K.; Zaccai, G. *Faraday Discuss.* **2009**, *141*, 117–130.

(26) Halle, B.; Andersson, T.; Forsen, S.; Lindman, B. *J. Am. Chem. Soc.* **1981**, *103*, 500–508.

(27) Polnaszek, C. F.; Bryant, R. G. *J. Chem. Phys.* **1984**, *81*, 4038–4045.

(28) Polnaszek, C. F.; Bryant, R. G. *J. Am. Chem. Soc.* **1984**, *106*, 428–429.

(29) Polnaszek, C. F.; Hanggi, D. A.; Carr, P. W.; Bryant, R. G. *Anal. Chim. Acta* **1987**, *194*, 311–315.

(30) Carlstrom, G.; Halle, B. *Langmuir* **1988**, *4*, 1346–1352.

(31) Armstrong, B. D.; Han, S. *J. Am. Chem. Soc.* **2009**, *131*, 4641–4647.

(32) Pal, S. K.; Peon, J.; Zewail, A. H. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 1763–1768.

(33) Rezus, Y. L. A.; Bakker, H. J. *J. Phys. Chem. A* **2008**, *112*, 2355–2361.

(34) Lee, S. H.; Rossky, P. J. *J. Chem. Phys.* **1994**, *100*, 3334–3345.

(35) Bizzarri, A. R.; Cannistraro, S. *Phys. Rev. E* **1996**, *53*, R3040–R3043.

(36) Rocchi, C.; Bizzarri, A.; Cannistraro, S. *Phys. Rev. E* **1998**, *57*, 3315–3325.

(37) Bizzarri, A. R.; Cannistraro, S. *J. Phys. Chem. B* **2002**, *106*, 6617–6633.

(38) Marchi, M.; Sterpone, F.; Ceccarelli, M. *J. Am. Chem. Soc.* **2002**, *124*, 6787–6791.

(39) Argyris, D.; Tummala, N. R.; Striolo, A.; Cole, D. R. *J. Phys. Chem. C* **2008**, *112*, 13587–13599.

(40) Johnson, M. E.; Malardier-Jugroot, C.; Murarka, R. K.; Head-Gordon, T. *The J. Phys. Chem. B* **2008**, *113*, 4082–4092.

(41) MacKerell, A. D.; et al. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.

(42) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1982**, *4*, 187–217.

(43) MacKerell, A.; Kuczera, J. W.; Karplus, M. *J. Am. Chem. Soc.* **1995**, *117*, 11946–11975.

(44) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatharm, T. E., III; De Bolt, S.; Ferguson, D.; Seibel, G.; Kollmann, P. *Comput. Phys. Commun.* **1995**, *91*, 1–41.

(45) Scott, W. R. P.; Hünenberg, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennen, J.; Torda, A. E.; Huber, T.; Krüger, P.; van Gunsteren, W. F. *J. Phys. Chem. A* **1999**, *103*, 3596–3607.

(46) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

(47) Guillot, B. *J. Mol. Liq.* **2002**, *101*, 219–260.

(48) Nutt, D. R.; Smith, J. C. *J. Am. Chem. Soc.* **2008**, *130*, 13066–13073.

(49) Cui, J.; Battle, K.; Wierzbicki, A.; Madura, J. D. *Int. J. Quantum Chem.* **2009**, *109*, 73−80; 16th Conference on Current Trends in Computational Chemistry, Jackson, MS, Nov. 2−3, 2007.

(50) Mahoney, M. W.; Jorgensen, W. L. *J. Chem. Phys.* **2000**, *112*, 8910–8922.

(51) Nandi, N.; Bagchi, B. *J. Phys. Chem. B* **1997**, *101*, 10954–10961.

(52) Vega, C.; Abascal, J. L. F.; Conde, M. M.; Aragones, J. L. *Faraday Discuss.* **2009**, *141*, 251–276.

(53) Kwac, K.; Lee, K. K.; Han, J. B.; Cho, M. *J. Chem. Phys.* **2008**, *128*, 105106.

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

(55) van der Spoel, D.; Lindahl, E. *J. Phys. Chem. B* **2003**, *107*, 11178–11187.

(56) Mark, P.; Nilsson, L. *J. Phys. Chem. B* **2001**, *105*, 8028–8035.

(57) Nutt, D. R.; Smith, J. C. *J. Chem. Theory Comput.* **2007**, *3*, 1550–1560.

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

(59) Parak, F. G.; Formanek, H. *Acta Crystallogr., Sect. A* **1971**, *27*, 573–578.

(60) Knapp, E. W.; Fischer, S. F.; Parak, F. *J. Phys. Chem.* **1982**, *86*, 5042–5047.

(61) Parak, F.; Knapp, E. W. *Proc. Natl. Acad. Sci. U.S.A.* **1984**, *81*, 7088–7092.

(62) Doster, W.; Cusack, S.; Petry, W. *Nature* **1989**, *337*, 754–756.

(63) Smith, J. C. *Q. Rev. Biophys.* **1991**, *24*, 227–291.

(64) Rasmussen, B. F.; Stock, A. M.; Ringe, D.; Petsko, G. A. *Nature* **1992**, *357*, 423–424.

(65) Fitter, J.; Lechner, R. E.; Dencher, N. A. *Biophys. J.* **1997**, *73*, 2126–2137.

(66) Ostermann, A.; Waschipky, R.; Parak, F. G.; Nienhaus, G. U. *Nature* **2000**, *404*, 205–208.

(67) Ringe, D.; Petsko, G. A. *Biophys. Chem.* **2003**, *105*, 667–680.

(68) Tournier, A. L.; Xu, J.; Smith, J. C. *Biophys. J.* **2003**, *85*, 1871–1875.

(69) Tournier, A. L.; Smith, J. C. *Phys. Rev. Lett.* **2003**, *91*, 208106.

(70) Schulz, R.; Krishnan, M.; Daidone, I.; Smith, J. C. *Biophys. J.* **2009**, *96*, 476–484.

(71) Roh, J. H.; Curtis, J. E.; Azzam, S.; Novikov, V. N.; Peral, I.; Chowdhuri, Z.; Gregory, R. B.; Sokolov, A. P. *Biophys. J.* **2006**, *91*, 2573–2588.

(72) Iben, I. E. T.; Braunstein, D.; Doster, W.; Frauenfelder, H.; Hong, M. K.; Johnson, J. B.; Luck, S.; Ormos, P.; Schulte, A.; Steinbach, P. J.; Xie, A. H.; Young, R. D. *Phys. Rev. Lett.* **1989**, *62*, 1916–1919.

(73) Ferrand, M.; Dianoux, A. J.; Petry, W.; Zaccai, G. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 9668–9672.

(74) Fitter, J.; Ernst, O. P.; Hauss, T.; Lechner, R. E.; Hofmann, K. P.; Dencher, N. A. *Eur. Biophys. J.* **1998**, *27*, 638–645.

(75) Fitter, J.; Verclasc, S. A. W.; Lechnerc, R. E.; Seelerta, H.; Dencher, N. A. *FEBS Lett.* **1998**, *433*, 321–325.

(76) Fitter, J. *Biophys. J.* **1999**, *76*, 1034–1042.

(77) Vitkup, D.; Ringe, D.; Petsko, G. A.; Karplus, M. *Nat. Struct. Mol. Biol.* **2000**, *7*, 34–38.

(78) Reat, V.; Dunn, R.; Ferrand, M.; Finney, J. L.; Daniel, R. M.; Smith, J. C. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 9961–9966.

(79) Fenimore, P. W.; Frauenfelder, H.; McMahon, B. H.; Parak, F. G. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 16047–16051.

(80) Paciaroni, A.; Cinelli, S.; Onori, G. *Biophys. J.* **2002**, *83*, 1157–1164.

(81) Chaplin, M. F. *Biochem. Mol. Biol. Edu.* **2001**, *29*, 54–59.

(82) Lee, A. L.; Wand, A. J. *Nature* **2001**, *411*, 501–504.

(83) Hayward, J. A.; Smith, J. C. *Biophys. J.* **2002**, *82*, 1216–1225.

(84) Roh, J. H.; Novikov, V. N.; Gregory, R. B.; Curtis, J. E.; Chowdhuri, Z.; Sokolov, A. P. *Phys. Rev. Lett.* **2005**, *95*, 038101.

(85) Doster, W.; Settles, M. *Biochim. Biophys. Acta* **2005**, *1749*, 173–186.

(86) Krishnan, M.; Kurkal-Siebert, V.; Smith, J. C. *J. Phys. Chem. B* **2008**, *112*, 5522–5533.

(87) Cordone, L.; Ferrand, M.; Vitrano, E.; Zaccai, G. *Biophys. J.* **1999**, *76*, 1043–1047.

(88) Tsai, A. M.; Neumann, D. A.; Bell, L. N. *Biophys. J.* **2000**, *79*, 2728–2732.

(89) Tarek, M.; Tobias, D. J. *Phys. Rev. Lett.* **2002**, *88*, 138101.

(90) Ding, X.; Rasmussen, B. F.; Petsko, G. A.; Ringe, D. *Biochemistry* **1994**, *33*, 9285–9293.

(91) Parak, F. G.; Frolov, E. N.; Kononenko, A. A.; Mössbauer, R. L.; Goldanskii, V. I.; Rubin, A. B. *FEBS Lett.* **1980**, *117*, 368–372.

(92) Daniel, R. M.; Smith, J. C.; Ferrand, M.; Hery, S.; Dunn, R.; Finney, J. L. *Biophys. J.* **1998**, *75*, 2504–2507.

(93) Dunn, R.; Reat, V.; Finney, J. L.; Ferrand, M.; Smith, J. C.; Daniel, R. M. *Biochem. J.* **2000**, *346*, 355–358.

(94) Bragger, J. M.; Dunn, R. V.; Daniel, R. M. *Biochim. Biophys. Acta* **2000**, *1480*, 278–282.

(95) Kurkal, V.; Daniel, R. M.; Finney, J. L.; Tehei, M.; Dunn, R. V.; Smith, J. C. *Biophys. J.* **2005**, *89*, 1282–1287.

(96) Neria, E.; Fischer, S.; Karplus, M. *J. Chem. Phys.* **1996**, *105*, 1902–1921.

(97) Mark, P.; Nilsson, L. *J. Phys. Chem. A* **2001**, *105*, 9954–9960.

(98) Vojtechovsky, J.; Chu, K.; Berendzen, J.; Sweet, R. M.; Schlichting, I. *Biophys. J.* **1999**, *77*, 2153–2174.

(99) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comp. Phys.* **1977**, *23*, 321–341.

(100) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(101) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

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

(103) Wong, C. F.; Zheng, C.; Mccammon, J. A. *Chem. Phys. Lett.* **1989**, *154*, 151–154.

(104) Arcangeli, C.; Bizzarri, A. R.; Cannistraro, S. *Chem. Phys. Lett.* **1998**, *291*, 7–14.

(105) Parak, F. G.; Knapp, E. W.; Kucheida, D. *J. Mol. Biol.* **1982**, *161*, 177–194.

(106) Tarek, M.; Tobias, D. J. *Biophys. J.* **2000**, *79*, 3244–3257.

(107) Steinbach, P.; Brooks, B. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 9135–9139.

(108) Teeter, M. M.; Yamano, A.; Stec, B.; Mohanty, U. *Proc. Natl. Acad. Sci. U. S. A.* **2001**, *98*, 11242.

(109) Caliskan, G.; Kisliuk, A.; Sokolov, A. P. *J. Non-Cryst. Solids* **2002**, *307−310*, 868–873.

(110) van der Spoel, D.; van Maaren, P. J. *J. Chem. Theory Comput.* **2006**, *2*, 1–11.

(111) Murarka, R. K.; Head-Gordon, T. *J. Chem. Phys.* **2007**, *126*, 215101.

(112) Rault, J. *J. Non-Cryst. Solids* **2000**, *271*, 177–217.

(113) Debenedetti, P. G.; Stillinger, F. H. *Nature* **2001**, *410*, 259–267.

(114) Kauzmann, W. *Chem. Rev.* **1948**, *43*, 219–256.

(115) Angell, C. A. *Science* **1995**, *267*, 1924–1935.

(116) Kirkwood, J. G. *J. Chem. Phys.* **1939**, *7*, 911–919.

(117) Boresch, S.; Ringhofer, S.; Höechtl, P.; Steinhauser, O. *Biophys. Chem.* **1999**, *78*, 43–68.

(118) Höechtl, P.; Boresch, S.; Bitomsky, W.; Steinhauser, O. *J. Chem. Phys.* **1998**, *109*, 4927–4937.

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

(120) Hermans, J.; Berendsen, H. J. C.; van Gunsteren, W. F.; Postma, J. P. M. *Biopolymers* **1984**, *23*, 1513–1518.

(121) Rezus, Y. L. A.; Bakker, H. J. *J. Chem. Phys.* **2005**, *123*, 114502.

(122) R$\phi$nne, C.; Thrane, L.; Åstrand, P. O.; Wallqvist, A.; Mikkelsen, K. V.; Keiding, S. R. *J. Chem. Phys.* **1997**, *107*, 5319–5331.

(123) Laage, D.; Hynes, J. T. *Science* **2006**, *311*, 832–835.

(124) McCammon, J. A.; Gelin, B. R.; Karplus, M. *Nature* **1977**, *267*, 585–590.

(125) Karplus, M.; Kushick, J. N. *Macromolecules* **1981**, *14*, 325–332.

(126) Smith, J. C.; Kuczera, K.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 1601–1605.

(127) Setny, P.; Wang, Z.; Cheng, L. T.; Li, B.; McCammon, J. A.; Dzubiella, J. *Phys. Rev. Lett.* **2009**, *103*, 187801.

CT9006508

JCTC Journal of Chemical Theory and Computation

# Constant pH Replica Exchange Molecular Dynamics in Biomolecules Using a Discrete Protonation Model

Yilin Meng and Adrian E. Roitberg*

*Department of Chemistry and Quantum Theory Project, University of Florida, Gainesville, Florida 32611-8435*

Received December 16, 2009

**Abstract:** A constant pH replica exchange molecular dynamics (REMD) method is proposed and implemented to improve coupled protonation and conformational state sampling. By mixing conformational sampling at constant pH (with discrete protonation states) with a temperature ladder, this method avoids conformational trapping. Our method was tested and applied to seven different biological systems. The constant pH REMD not only predicted p$K_a$ correctly for small, model compounds but also converged faster than constant pH molecular dynamics (MD). We further tested our constant pH REMD on a heptapeptide from the ovomucoid third domain (OMTKY3). Although constant pH REMD and MD produced very close p$K_a$ values, the constant pH REMD showed its advantage in the efficiency of conformational and protonation state samplings.

## Introduction

Solution pH is a very important thermodynamic variable that affects protein structure, function, and dynamics.[1−3] Many biological phenomena such as protein folding/misfolding,[4−6] substrate docking,[7,8] and enzyme catalysis[9−11] are pH-dependent. Examples include amyloid fibril formation[12] such as misassembly of prion proteins,[13] ATP synthesis,[14] and pH-dependent partial α-helical formation of a 13-residue N-terminal fragment from ribonuclease A.[4,5] This pH-dependence of structure and dynamics comes from changes in the ratio of protonation states for the different residues at different solution pH values.

The pH value at which a particular titratable residue side chain has equal population of protonated and deprotonated states is called the p$K_a$ value of that side chain.[15−18] The p$K_a$ value of a titratable side chain can be highly affected by the environment of that titratable side chain such as protein environment polarity. An ionizable side chain in the interior of a protein can have a different p$K_a$ value from the isolated amino acid in solution.[18] For example, Asp26 of thioredoxin, which lies in a deep pocket of the protein, has a p$K_a$ value of 7.5, while the intrinsic p$K_a$ value of aspartic acid is 4.0.[19] Furthermore, a charged side chain can favor different protonation states in order to stabilize the protein structure by forming a salt bridge.[20] The conformation and protonation distributions are highly coupled:[21−23] changes in either of them can affect the other one.

Due to the importance of solution pH, Molecular Dynamics (MD) simulations have been used to study its effect on protein structure and dynamics. Other popular theoretical methods developed to calculate (predict) p$K_a$ values include the electrostatic continuum dielectric model and the Poisson−Boltzmann Equation (PBE),[17,24−27] free energy calculation methods,[16,28−30] and empirical methods.[31,32] More details on computer simulation of p$K_a$ prediction and pH dependence of protein structure and dynamics can be found in recent studies.[33−51] The traditional way of studying the effect of pH is setting a constant protonation state before a simulation is carried out. The major problem with this method is that it decouples the correlation between conformation and protonation state, yielding a wrong population of protonation states, especially when the solution pH is close to the p$K_a$ of that titratable site. Furthermore, assigning protonation states before a simulation often involves a guess of protonation state based on our experience.

Constant-pH molecular dynamics (constant-pH MD) methods were developed in order to correlate the protein conformation and protonation state. The purpose of constant-pH MD is to describe protonation equilibrium correctly at a given pH. One category of constant-pH MD methods uses a

* Corresponding author e-mail: roitberg@ufl.edu.

10.1021/ct900676b © 2010 American Chemical Society
Published on Web 03/10/2010

continuous protonation parameter. Earlier models include a grand canonical MD algorithm developed by Mertz and Pettitt[52] in 1994 and a method introduced by Baptista et al.[35] in 1997. In the Mertz and Pettitt model, protons are allowed to be exchanged between a titratable side chain and water molecules. Baptista et al. used a potential of mean force to treat protonation and conformation simultaneously. Later, Börjesson and Hünenberger[53,54] developed a continuous protonation variable model in which the protonation fraction is adjusted by weak coupling to a proton bath, using an explicit solvent. More recently, the continuous protonation state model was further developed by the Brooks group.[39−43,55] They called their constant-pH MD algorithm continuous constant-pH molecular dynamics (CPHMD). In the CPHMD method, Lee et al.[55] applied $\lambda$-dynamics[56] to the protonation coordinate and used the Generalized Born (GB) implicit solvent model. They chose a $\lambda$ variable to control the protonation fraction and introduced an artificial potential barrier between protonated and deprotonated states. The potential is a biasing potential to increase the residency time close to protonation/deprotonation states, and it centered at the half-way point of titration ($\lambda = 1/2$). The CPHMD method was then extended by incorporating an improved GB model and replica exchange molecular dynamics (REMD) algorithm for better sampling.[40−43] The applications of CPHMD and replica exchange CPHMD included predicting p$K_a$ values of various proteins,[40] studying proton tautomerism[39] and pH-dependent protein folding and folding intermediates of the villin headpiece domain.[42,43]

In addition to continuous protonation state models, discrete protonation state methods have also been developed to study the pH dependence of protein structure and dynamics.[36,46−49,57−63] The discrete protonation state models utilize a hybrid molecular dynamics and Monte Carlo (hybrid MD/MC) method. Protein conformations are sampled by molecular dynamics, and protonation states are sampled using a Monte Carlo scheme periodically during a MD simulation. A new protonation state is selected after a user-defined number of MD steps, and the free energy difference between the old and the new state is calculated. The Metropolis criterion[64] is used to accept or reject the protonation change. Various solvent models and protonation state energy algorithms were used in discrete protonation state constant pH MD simulations. The Baptista group[36,46−49] used the Poisson−Boltzmann (PB) equation to calculate protonation energies while their MD was done in explicit solvent. Walczak and Antosiewicz[63] also employed the PB equation to determine protonation energy, but they used Langevin dynamics to propagate coordinates between MC steps. Bürgi et al.[57] calculated the transition energy between two protonation states by using thermodynamic integration (TI) method and explicit solvent. More recently, Mongan et al.[62] developed a method combining the GB model[65,66] and the discrete protonation state model. In Mongan's method, the GB model was used in protonation state transition energy as well as solvation free energy calculations. Therefore, solvent models in conformational and protonation state sampling are consistent, and the computational cost is small. This model was later coupled with accelerated molecular dynamics[67,68] to achieve better conformational sampling.[69] Dlugosz and Antosiewicz also used the discrete protonation state method to study succinic acid[58] and a heptapeptide derived from the ovomucoid third domain (OMTKY3).[60,61] This heptapeptide corresponds to residue 26−32 of OMTKY3 and has the sequence of acetyl−Ser−Asp−Asn−Lys−Thr−Tyr−Gly−methylamine. Nuclear magnetic resonance (NMR) experiments indicated the p$K_a$ of Asp is 3.6, 0.4 p$K_a$ unit lower than the value of blocked Asp dipeptide.[61] In their studies, the conventional molecular dynamics (MD) simulations were carried out to sample peptide conformations. Dlugosz and Antosiewicz sampled protonation states using the PB equation and used analytical continuum electrostatics to treat solvation effects. Their method predicted the p$K_a$ to be 4.24.

Due to the correlation between conformation and protonation sampling, correct sampling of protonation states requires accurate sampling of protein conformations. Hence, generalized ensemble methods[70−73] such as the multicanonical ensemble algorithm,[74,75] simulated tempering,[76] and replica exchange molecular dynamics (REMD)[77] should be used to avoid kinetic trapping which comes from low rates of barrier crossing in constant temperature MD simulations. These methods make the system perform a random walk in temperature or energy space which allows the system under study to easily overcome energy barriers and hence reduces the problem of kinetic trapping. REMD, the MD version of parallel tempering (PT),[78] has the advantage of a-priori known weight factors, such as Boltzmann weights. REMD has been used in many studies of protein structure and dynamics and proven to drastically increase rates of convergence toward a proper equilibrium distribution. Khandogin et al. applied the REMD algorithm to the continuous protonation state constant-pH method and named it REX-CPHMD. They applied REX-CPHMD to p$K_a$ predictions and pH-dependent protein dynamics such as folding and aggregations.[40−43]

In this paper, we present a study of conformation and protonation state sampling using an REMD algorithm on the discrete protonation state model proposed by Mongan et al. We first tested our method on the basis of five dipeptides and a model peptide having the sequence Ala−Asp−Phe−Asp−Ala (ADFDA). The two ends of model peptide ADFDA were not capped, so the two ionizable side chains would have different environments. Then our method was applied to a heptapeptide from OMTKY3, the same heptapeptide that Dlugosz and Antosiewicz studied in their paper.[60,61] Our purpose is to show that the REMD algorithm coupled with a discrete protonation state description can greatly improve pH-dependent protein conformation and protonation state sampling.

## Methods

**A. Constant-pH MD Algorithm in AMBER.** A detailed description of the discrete protonation state model can be found in the paper of Mongan et al.[62] This algorithm employs discrete protonation states, MC sampling of protonation states, and the use of a GB model in MD and MC. Given a protein with N titratable sites, the discrete protonation state model means protonation states of a protein are described by a vector $\mathbf{n} = (n_1, n_2, ..., n_N)$ where each $n_i$ is some integer representing the protonation state of titratable residue $i$. In AMBER, five amino acids are designed to be titratable: aspartate, glutamate, histidine, lysine, and tyrosine. For each titratable residue, different protonation states have different partial charges on the side chain. This model also includes syn and anti forms of protons for the aspartate and glutamate side chains as well as the $\delta$ and $\varepsilon$ proton locations for histidine.

The goal of constant-pH MD is to describe equilibrium between protonated and deprotonated forms correctly at a given pH. In the discrete protonation model, the populations of each form are sampled by the MC method periodically during a MD simulation. At each Monte Carlo step, a titratable site and a new protonation state for that site are chosen randomly, and the transition free energy at this fixed configuration is used to evaluate the MC move.

Considering a titratable site A in a protein environment, its protonated form is protA-H and the deprotonated form is protA$^-$. The equilibrium between the two forms is governed by their free energy difference. This free energy difference is the ensemble average of different configurations. However, the free energy difference cannot be computed by a molecular mechanics (MM) model since the transition between two forms deals with bond breaking/forming and solvation of a proton which involves quantum mechanical effects.

The above problems can be solved by using a reference compound. The reference compound has the same titratable side chain as protA-H but with a known p$K_a$ value (p$K_{a,ref}$). Following Mongan et al.,[62] we assume the transition free energy can be divided into the quantum mechanics (QM) part and the molecular mechanics (MM) part. We further assume that the quantum mechanical energy components are the same between the reference compound and the protA-H. Since the p$K_a$ of the reference compound is known, its transition free energy from the deprotonated form to the protonated form at a given pH is

$$\Delta G_{\mathrm{ref}} = k_{\mathrm{B}}T \ln 10(\mathrm{pH} - \mathrm{p}K_{\mathrm{a,ref}}) \tag{1}$$

So the QM component of the transition free energy can be expressed as

$$\Delta G_{\mathrm{ref,QM}} = \Delta G_{\mathrm{ref}} - \Delta G_{\mathrm{ref,MM}} \tag{2}$$

where $\Delta G_{\mathrm{ref,MM}}$ is the molecular mechanics contribution to the free energy of the protonation reaction for that reference compound. In practice, the QM component of the transition free energy also contains errors from MM calculations, so it is actually better called a non-MM component. Since the approximation of the QM component of the transition free energy is

$$\Delta G_{\mathrm{ref,QM}} = \Delta G_{\mathrm{protein,QM}} \tag{3}$$

then the transition free energy from protA$^-$ to protA-H can be calculated as

$$\Delta G = k_{\mathrm{B}}T \ln 10(\mathrm{pH} - \mathrm{p}K_{\mathrm{a,ref}}) + \Delta G_{\mathrm{MM}} - \Delta G_{\mathrm{ref,MM}} \tag{4}$$

where $\Delta G_{\mathrm{MM}}$ is the molecular mechanics contribution (electrostatic interactions in nature) to the free energy of the protein titratable site. Hence, by using a reference compound, the QM effects are not needed. Effectively, we compute $\Delta$p$K_a$ relative to the reference compound. Computing $\Delta$p$K_a$ can also help canceling some errors introduced by the GB solvation model through the use of $\Delta G_{\mathrm{ref,MM}}$. In AMBER, a reference compound is a blocked dipeptide amino acid possessing a titratable side chain (for example, acetyl−Asp−methylamine). Five reference compounds were constructed corresponding to five titratable residues. The values of $\Delta G_{\mathrm{ref,MM}}$ for each reference compound are obtained from thermodynamic integration calculations at 300 K and set as internal parameters in AMBER.[62,79] The $\Delta G_{\mathrm{MM}}$ is calculated by taking the difference between the potential energy with the charges of the current protonation state and the potential energy with the charges of the new protonation state (i.e., $\Delta G_{\mathrm{MM}}$ is approximately $\Delta H$ by averaging over configurations).

The $\Delta G$ from eq 4 is used to decide if a MC move in protonation space should be accepted or rejected. If the transition is accepted, MD steps are carried out to sample conformational space in the new protonation state. If the MC attempt is rejected, MD steps are also carried out with no change to the protonation state.

**B. Titration Curve and p$K_a$ Prediction Calculation.** The titration curve of an ideal titratable site having no interaction with other titratable groups follows the Henderson−Hasselbalch (HH) equation:

$$\mathrm{p}K_{\mathrm{a}} = \mathrm{pH} - \log\left(\frac{[\mathrm{A}^-]}{[\mathrm{HA}]}\right) \tag{5}$$

Molecular dynamics runs are assumed to be ergodic; thus the ratio of time that a titratable site spends in protonated and deprotonated states can be used as concentrations. The analytical form of the titration curve can be obtained by exponentiating both sides of the HH equation. A more generalized form of the HH equation which studies an ionizable residue interacting with another one can be written as

$$\mathrm{p}K_{\mathrm{a}} = \mathrm{pH} - n\log\left(\frac{[\mathrm{A}^-]}{[\mathrm{HA}]}\right) \tag{6}$$

So the titration curve of an interacting ionizable residue can be expressed as

$$s = \frac{1}{1 + 10^{n\cdot(\mathrm{p}K_{\mathrm{a}} - \mathrm{pH})}} \tag{7}$$

where s is the fraction of deprotonation and $n$ is the Hill coefficient. A Hill plot, which can be obtained by plotting log([A$^-$]/[HA]) as a function of pH, is used to study titration behavior. The HH equation (including its generalized form) will be represented as a straight line in a Hill plot. The $x$ intercept is the p$K_a$ value, and the slope is the Hill coefficient which reflects interactions between titratable residues.

**C. Replica Exchange Molecular Dynamics (REMD).** A detailed description of the REMD algorithm can be found in the papers of Sugita and Okamoto.[77] In REMD, $N$ noninteracting copies (replicas) of a system are simulated at

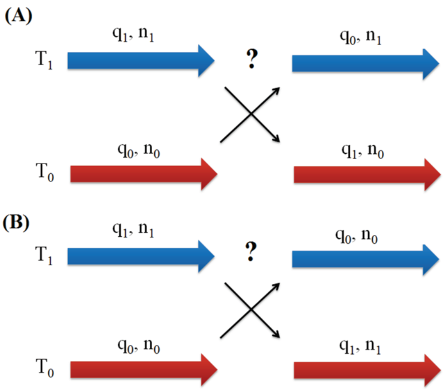


***Figure 1.*** Diagrams displaying exchanging algorithms in constant-pH REMD. (A) Only molecular structures (denoted as *q*) are attempted to exchange. In this case, protonation states are not touched at an exchange attempt. (B) Both molecular structures (denoted as *q*) and protonation states (denoted as *n*) are attempted to exchange at the same time. Metropolis criterion is applied in both algorithms to evaluate transitions.

*N* different temperatures (one each). Regular molecular dynamics is performed, and periodically an exchange of conformation between two (usually adjacent) temperatures is attempted. Suppose replica *i* at temperature $T_n$ and replica *j* at temperature $T_m$ are attempting to exchange; the following satisfies the detailed balance condition:

$$P_n(i)\,P_m(j)\,w(i \rightarrow j) = P_m(i)\,P_n(j)\,w(j \rightarrow i) \tag{8}$$

Here, $w(i \rightarrow j)$ is the transition probability between two states *i* and *j* and $P_n(i)$ is the population of state *i* at temperature *n* (in REMD assumed Boltzmann weighted). If the Metropolis criterion is applied, the exchange probability is obtained as

$$w(i \rightarrow j) = \min\{1, \exp[(\beta_m - \beta_n)(E(q^{[i]}) - E(q^{[j]}))]\} \tag{9}$$

Here, $q^{[i]}$ is the molecular configuration of state *i*, *E* is the potential energy, and $\beta = 1/k_BT$. If the exchange between two replicas is accepted, the temperatures of the two replicas will be swapped and velocities rescaled to the new temperatures by multiplying all the old velocities by the square root of the new temperature to old temperature ratio:

$$v_{\text{new}} = v_{\text{old}}\sqrt{\frac{T_{\text{new}}}{T_{\text{old}}}} \tag{10}$$

In the case of constant pH molecular dynamics, the potential energy of the system depends not only on the protein structure but also on the protein protonation state. Likewise, when coupling the REMD algorithm with constant-pH MD, one can either attempt to exchange molecular structures only or swap both structures and protonation states at the same time. For simplicity, let us consider two replicas where replica 0 has temperature $T_0$, protein structure $q_0$, and protonation state $n_0$, while replica 1 has temperature $T_1$, structure $q_1$, and protonation state $n_1$. A diagrammatic description of the two exchange algorithms is shown in Figure 1. The first way of performing an exchange attempt is that replica 0 tries to jump from state ($q_0$, $n_0$) to state ($q_1$, $n_0$) at temperature $T_0$ in one Monte Carlo step. Similarly, replica 1 attempts to transit from state ($q_1$, $n_1$) to state ($q_0$, $n_1$) at temperature $T_1$. Protonation states are kept at exchange attempts and only change during dynamics. Therefore, the detailed balance equation now becomes

$$\frac{w(\beta_0 q_0 n_0, \beta_1 q_1 n_1 \rightarrow \beta_0 q_1 n_0, \beta_1 q_0 n_1)}{w(\beta_0 q_1 n_0, \beta_1 q_0 n_1 \rightarrow \beta_0 q_0 n_0, \beta_1 q_1 n_1)} = \frac{\exp(-\beta_0 E(q_0, n_0))\exp(-\beta_1 E(q_1, n_1))}{\exp(-\beta_0 E(q_1, n_0))\exp(-\beta_1 E(q_0, n_1))} \tag{11}$$

Here, $w(\beta_0 q_0 n_0, \beta_1 q_1 n_1 \rightarrow \beta_0 q_1 n_0, \beta_1 q_0 n_1)$ is the transition probability of swapping structures. If the Metropolis criterion is used, this exchange probability can be written as

$$w(\beta_0 q_0 n_0, \beta_1 q_1 n_1 \rightarrow \beta_0 q_1 n_0, \beta_1 q_0 n_1) = \min\{1, \exp(-\Delta)\} \tag{12}$$

and

$$\Delta = \beta_0(E(q_0, n_0) - E(q_1, n_0)) - \beta_1(E(q_0, n_1) - E(q_1, n_1)) \tag{13}$$

where $\beta_0 = 1/k_BT_0$, $\beta_1 = 1/k_BT_1$, and *E* is the potential energy. Here, if the protonation states of two adjacent replicas at an exchange attempt are the same, the exchange probability of our constant pH REMD will be equivalent to the conventional REMD exchange probability. However, if it is not the case, four potential energy terms are needed to calculate exchange probability. Under this circumstance, the constant-pH REMD becomes a REMD algorithm that combines both temperature and Hamiltonian REMD algorithms.

One possible concern of exchanging only structures would be the role of kinetic energy, especially when $n_0$ and $n_1$ are different. In the REMD algorithm developed by Sugita and Okamoto, the kinetic energy terms in the Boltzmann factors cancel each other on average through velocity rescaling (eq 10). Only potential energies are required to compute exchange probabilities. There is a problem in canceling kinetic energy terms when the numbers of particles of two systems attempting to exchange are not the same. However, according to the constant-pH MD algorithm proposed by Mongan et al.,[62] a proton does not leave the molecule but becomes a dummy atom when an ionizable side chain is in a deprotonated state. Furthermore, that dummy atom retains its position and velocity, which are controlled by molecular dynamics. Hence, the kinetic energy contributions to the Boltzmann weight will be canceled out during exchange probability calculation, leaving only potential energy useful for the calculation.

The second possibility consists of exchanging protonation states as well as molecular structures at REMD Monte Carlo moves. For instance, replica 0 attempts to move from state ($q_0$, $n_0$) to state ($q_1$, $n_1$) at temperatures $T_0$ in one MC move, and replica 1 attempts to jump from state ($q_1$, $n_1$) to state ($q_0$, $n_0$) at temperature $T_1$. The detailed balance equation now can be written as

$$\frac{w(\beta_0 q_0 n_0, \beta_1 q_1 n_1 \rightarrow \beta_0 q_1 n_1, \beta_1 q_0 n_0)}{w(\beta_0 q_1 n_1, \beta_1 q_0 n_0 \rightarrow \beta_0 q_0 n_0, \beta_1 q_1 n_1)} = \frac{w(\beta_1 q_1 n_1 \rightarrow \beta_1 q_0 n_0)}{w(\beta_1 q_0 n_0 \rightarrow \beta_1 q_1 n_1)} \cdot \frac{w(\beta_0 q_0 n_0 \rightarrow \beta_0 q_1 n_1)}{w(\beta_0 q_1 n_1 \rightarrow \beta_0 q_0 n_0)} \tag{14}$$

***Table 1.*** The REMD p$K_a$ Predictions of Reference Compounds[a]

| p$K_a$ | aspartate | glutamate | histidine | lysine | tyrosine |
|---|---|---|---|---|---|
| REMD | 3.97 (0.01) | 4.41 (0.01) | 6.40 (0.03) | 10.42 (0.01) | 9.61 (0.01) |
| reference | 4.0 | 4.4 | 6.5 | 10.4 | 9.6 |

[a] The numbers in parentheses are the standard errors.

This equation states that the exchange probability is the product of MC transition probabilities at temperatures $T_0$ and $T_1$. If the protonation states of two adjacent replicas are the same at an exchange attempt, the exchange probability of constant-pH REMD becomes the exchange probability of conventional temperature-based REMD. If $n_0$ and $n_1$ are different, then each MC transition is essentially the protonation state change step in constant-pH MD, plus a structural transition. For example, consider the MC transition at temperature $T_0$

$$w(\beta_0 q_0 n_0 \rightarrow \beta_0 q_1 n_1) = \min\{1, \exp(-\Delta_1)\} \qquad (15)$$

where

$$\Delta_1 = \beta_0[E(q_1, n_0) - E(q_0, n_0)] + (\mathrm{pH} - \mathrm{p}K_{\mathrm{a,ref}}) + \beta_0[E_{\mathrm{elec}}(q_1, n_1) - E_{\mathrm{elec}}(q_1, n_0)] - \beta_0 \Delta G_{\mathrm{ref,MM}} \qquad (16)$$

The first term in $\Delta_1$ derives from the transition in configuration at fixed protonation state $n_0$, and the rest corresponds to protonation state change at fixed structure $q_1$. $E_{\mathrm{elec}}$ represents the electrostatic component of potential energy. Similarly, the transition probability of a MC jump at $T_1$ can be expressed as

$$w(\beta_1 q_1 n_1 \rightarrow \beta_1 q_0 n_0) = \min\{1, \exp(-\Delta_2)\} \qquad (17)$$

where

$$\Delta_2 = \beta_1[E(q_0, n_1) - E(q_1, n_1)] - (\mathrm{pH} - \mathrm{p}K_{\mathrm{a,ref}}) - \beta_1[E_{\mathrm{elec}}(q_0, n_1) - E_{\mathrm{elec}}(q_0, n_0)] + \beta_1 \Delta G_{\mathrm{ref,MM}} \qquad (18)$$

Therefore, according to eq 14, the exchange probability can be written as

$$w(\beta_0 q_0 n_0, \beta_1 q_1 n_1 \rightarrow \beta_0 q_1 n_1, \beta_1 q_0 n_0) = \min\{1, \exp(-\Delta')\} \qquad (19)$$

and

$$\Delta' = \Delta + \beta_0[E_{\mathrm{elec}}(q_1, n_1) - E_{\mathrm{elec}}(q_1, n_0)] - \beta_1[E_{\mathrm{elec}}(q_0, n_1) - E_{\mathrm{elec}}(q_0, n_0)] + (\beta_0 - \beta_1)\cdot \Delta G_{\mathrm{ref,MM}} \qquad (20)$$

where $\Delta$ is the same quantity as in eq 13.

The exchange probability calculation in the second method of coupling REMD and constant-pH MD utilizes the same number of energy evaluations required by the first method since obtaining electrostatic potential energies does not require extra energy calculations. The advantage of implementing the second exchanging protocol (exchange both structures and protonation states) over the first one (exchange structures only) should not be significant because it is the conformational sampling at higher temperature that greatly improves conformational sampling at lower temperatures. Allowing protonation states to change at exchange attempts does not provide extra gains in conformational sampling since the protonation state space is well sampled during the MD propagation. Therefore, only the first method of performing exchanges was implemented.

**D. Simulation Details.** For our study, constant pH REMD simulations were carried out first on five reference compounds: blocked aspartate, glutamate, histidine, lysine, and tyrosine, to test our method and implementation. The experimental p$K_a$ values of those reference compounds are known[80] and listed in Table 1. We later performed constant pH REMD simulations on a model peptide ADFDA (Ala−Asp−Phe−Asp−Ala, unblocked termini) and a heptapeptide derived from OMTKY3 (residues 26 to 32 with blocked termini). Four replicas were used in the reference compounds and ADFDA REMD simulations. The temperatures were 240, 300, 370, and 460 K for all six molecules. The pH range for the study of acidic side chains was sampled from 2.5 to 6, and the pH range of histidine is from 5.5 to 8. The basic side chains were titrated from pH 9 to 12. An interval of 0.5 was chosen for all titrations.

Eight replicas were chosen for the heptapeptide with a temperature range from 250 to 480 K. A total of 10 ns was used for each replica in all REMD simulations, and an exchange was attempted every 2 ps. A MC move to change the protonation state was attempted every 10 fs. A second set of REMD runs was done with the same overall conditions but different initial structures in order to check simulation convergence.

To compare conformational and protonation state sampling, 100 ns of constant pH MD simulations were carried out for the aspartate reference compound and ADFDA, at the same pH values as in the REMD runs. For the heptapeptide, one set of 10 ns constant pH MD simulations was done at each pH value simulated by the REMD method.

Constant pH REMD and MD simulations were done using the AMBER 10 molecular simulation suite.[81] The AMBER ff99SB force field[82] was used in all the simulations. The SHAKE algorithm[83] was used to constrain the bonds connecting hydrogen atoms in all the simulations which allowed use of a 2 fs time step. The OBC generalized Born implicit solvent model[66] was used to model the water environment in all our calculations. The Berendsen thermostat,[84] with a relaxation time of 2 ps, was used to keep the replica temperatures around their target values. The salt concentration (Debye−Huckel based) was set at 0.1 M. The cutoff for nonbonded interaction and the Born radii was 30 Å.

**E. Cluster Analysis.** Cluster analysis was done using the Moil-View program[85] in order to compare conformational sampling.[86,87] The MD and REMD trajectories (having same number of frames) at 300 K and under the same solvent pH were combined following a procedure introduced in the paper of Okur et al. Then, the combined trajectory was clustered on the basis of peptide backbone atoms' root-mean-square deviations (RMSDs). The population fraction corresponding to each cluster was obtained for MD and REMD simulation. The correlation coefficient values which represent the correlations between MD and REMD cluster population were calculated at each solution pH value by doing linear regression. A high correlation between MD and REMD cluster populations indicates that the structure ensembles are close to each other. This method provides a direct comparison of global conformational sampling between MD and REMD simulations. The same technique was used when studying the convergence of constant pH REMD and MD trajectories. A cluster cutoff RMSD of 1.5 Å is chosen for both ADFDA and the heptapeptide during our analysis.

**F. Local Conformational Sampling and Its Convergence to the Final State.** In our study, the local conformational sampling was examined by comparing the probability density of the backbone dihedral angle pair ($\varphi$, $\Psi$). Essentially, we are comparing the Ramachandran plot of a residue. Each ($\varphi$, $\Psi$) probability density was computed by binning $\varphi$ and $\Psi$ angle pairs 10° × 10°. These two-dimensional histograms were normalized into populations, and the contours were plotted. The metric used to evaluate ($\varphi$, $\Psi$) probability density convergence was the root-mean-squared deviation (RMSD) between the cumulative ($\varphi$, $\Psi$) histogram and the one produced by using all configurations. Each cumulative histogram was constructed by using ($\varphi$, $\Psi$) pairs up to the current time and following the same algorithm mentioned earlier in this section.

## Results and Discussion

**A. Reference Compounds.** We first applied our constant pH REMD method to the reference compounds. Table 1 shows the p$K_a$ values predicted by REMD simulations (10 ns for each replica) as well as the reference p$K_a$ values. All our p$K_a$ values were calculated by fitting to the HH equation. Agreement between constant pH REMD predictions and the reference values can be seen.

The pH titration curves of the same reference compounds showed agreement between MD (100 ns) and REMD simulations. Figure 2 demonstrates the REMD and MD titration curves of aspartic acid reference compound as an example.

We further studied the convergence of protonation states sampling. REMD and MD protonation fractions were plotted with respect to MC attempts for the aspartate reference compound at all pH values. Figure 3A demonstrates the protonated fraction versus time at pH 4 as one example. According to Figure 3A, it suggests that, although the final p$K_a$ predictions are the same for REMD and MD simulations, the protonation state sampling during REMD simulations clearly converges faster than that in a MD run.

**B. Model Peptide ADFDA.** The model peptide ADFDA (as zwitterion) was chosen as a more stringent test of our constant pH REMD method. The charged termini will provide a different electrostatic environment for each titratable Asp residue, and hence a correct constant pH REMD model should reflect this difference between titration curves of the two Asp residues. The Asp2 residue is closer to the $NH_3^+$, so the deprotonated state is favored, and the p$K_a$ value of Asp2 residue should shift below 4.0 (which is the p$K_a$ of the reference aspartic dipeptide). The Asp4 residue is closer

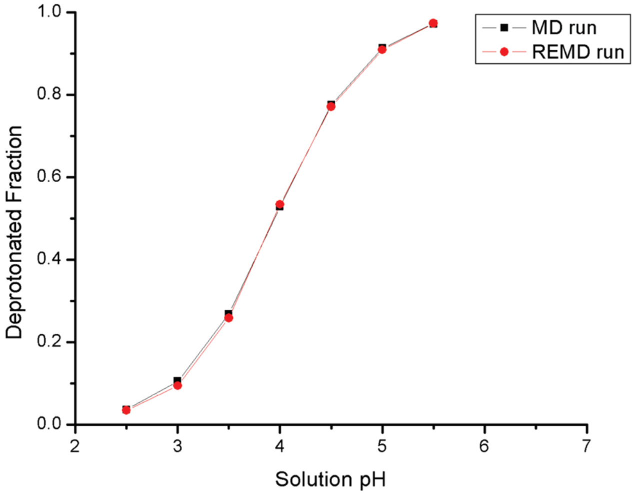


***Figure 2.*** Titration curves of blocked aspartate amino acid from 100 ns MD at 300 K and REMD runs. Agreement can be seen between MD and REMD simulations.

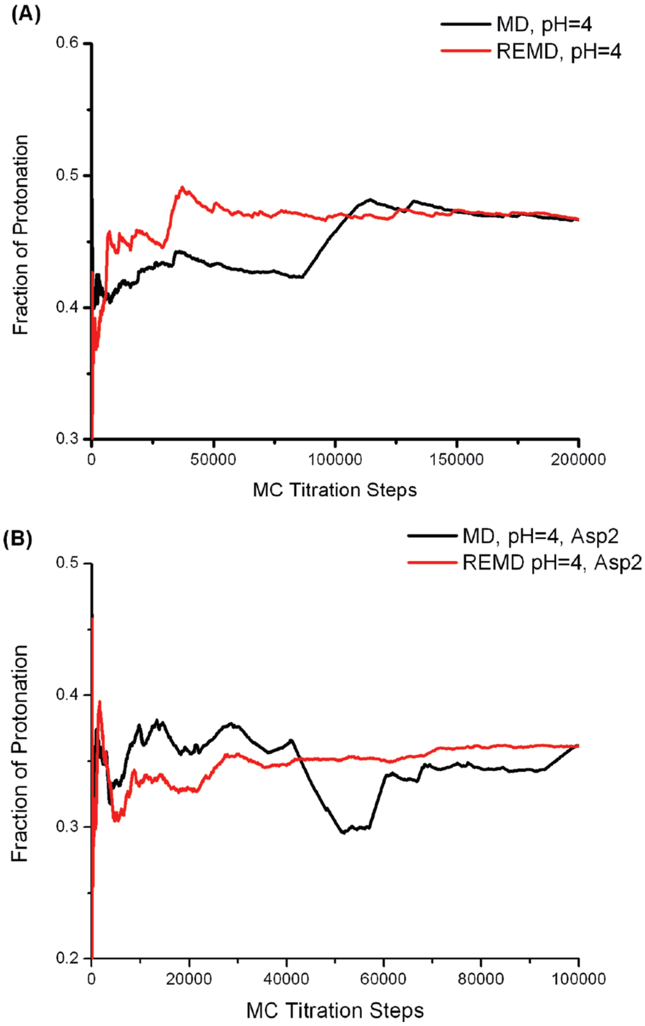


***Figure 3.*** Cumulative average protonation fraction of a titratable residue vs Monte Carlo (MC) steps. (A) Aspartic acid reference compound at pH = 4. (B) Asp2 in model peptide ADFDA at pH = 4.

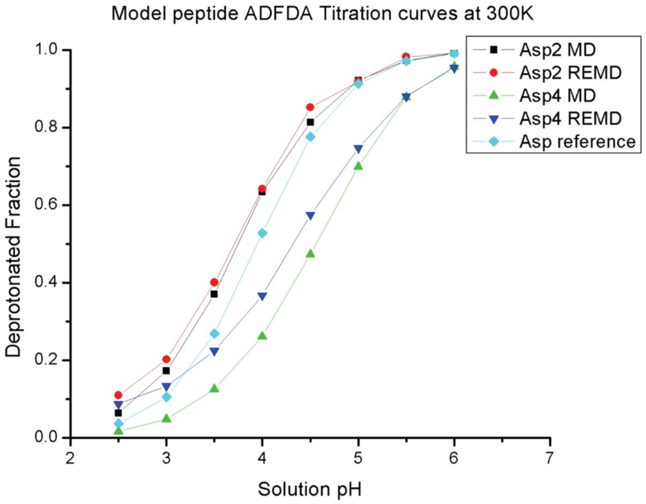


***Figure 4.*** The titration curves of the model peptide ADFDA at 300 K from both MD and REMD simulations. MD simulation time was 100, and 10 ns were chosen for each replica for REMD runs.

***Table 2.*** p$K_a$ Predictions and Hill Coefficients Fitted from the HH Equation

| | Asp2 | | Asp4 | |
|---|---|---|---|---|
| | p$K_a$ | Hill coefficient | p$K_a$ | Hill coefficient |
| REMD | 3.74 | 0.87 | 4.38 | 0.67 |
| MD | 3.76 | 0.89 | 4.54 | 0.85 |

to the COO$^-$ negative charge, and hence the p$K_a$ value should shift above 4.0.

The titration curves of the model peptide ADFDA from REMD simulations are shown in Figure 4. We can clearly see that Asp2 and Asp4 have different titration curves from each other and from the reference compound. The p$K_a$ value and Hill coefficient for each Asp residue were obtained by fitting titration curves to a Hill plot. The results are shown in Table 2. The REMD p$K_a$ predictions reflect the difference between Asp2 and Asp4 due to different peptide electrostatic environments. We also displayed the MD titration curves of Asp2 and Asp4 in Figure 4 and listed the MD p$K_a$ predictions and corresponding Hill coefficients in Table 2. The titration curve of the Asp2 residue only showed a small difference between MD and REMD simulations. But we can see differences in titration behaviors of Asp4 between MD and REMD calculations when the solution pH is below 5. Interestingly, Lee et al. studied blocked Asp−Asp peptide using the CPHMD method,[55] reporting different Hill coefficients for each of the two Asp residues.

Convergence rates of Asp2 titration behavior were compared between REMD and MD calculations due to the fact that Asp2 titration curves are very close. The cumulative protonated fractions versus MC attempts at pH 4 are shown in Figure 3B. Faster convergence in protonation state sampling can be seen for REMD simulation even though both REMD and MD calculations resulted in the same final protonated fraction. Clearly, our constant pH REMD method accelerates the convergence of sampling of protonation states.

In addition to protonation state sampling, we also evaluated the conformational sampling in constant pH MD and REMD simulations. First, distributions of backbone $\varphi$ and $\Psi$ angle pairs (Ramachandran plots) of residues Asp2, Phe3, and Asp4 in ADFDA at each solution pH were studied. The regions in Ramachandran plots sampled by MD and REMD simulations are the same. Ramachandran plots for residue Asp2 at pH 4 are shown in Figure 5 as an example.

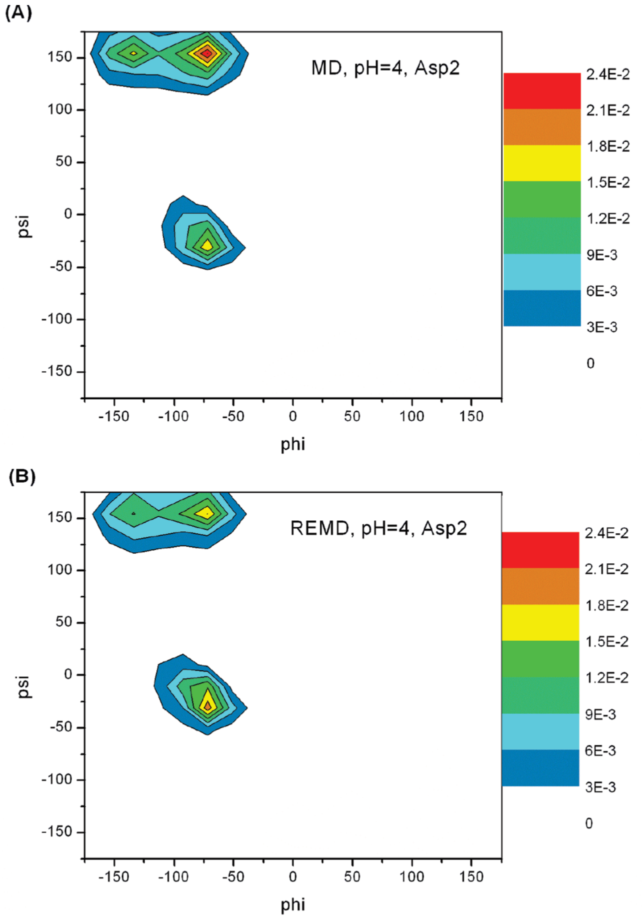


***Figure 5.*** Backbone dihedral angle ($\varphi$, $\Psi$) normalized probability density (Ramachandran plots) for Asp2 at pH 4 in ADFDA. Ramachandran plots at other solution pH values are similar. For Asp2, constant-pH MD and REMD sampled the same local backbone conformational space. Phe3 and Asp4 Ramachandran plots also display the same trend.

***Table 3.*** Correlation Coefficient between MD and REMD Cluster Populations[a]

| | pH = 2.5 | pH = 3 | pH = 3.5 | pH = 4 |
|---|---|---|---|---|
| $R^2$ | 0.94 | 0.90 | 0.79 | 0.93 |
| | pH = 4.5 | pH = 5 | pH = 5.5 | pH = 6 |
| $R^2$ | 0.85 | 0.98 | 0.92 | 0.96 |

[a] The $R^2$ values were calculated by linear regression.

Since the Ramachandran plots only represent local conformational sampling, we also evaluated global conformational sampling by clustering MD and REMD trajectories and comparing the cluster populations. The MD and REMD cluster population $R^2$ values are listed in Table 3. A plot of cluster populations from MD and REMD trajectories at a solution pH of 4 is shown in Figure 6A as an example. The large $R^2$ values indicate that the MD and REMD sampled

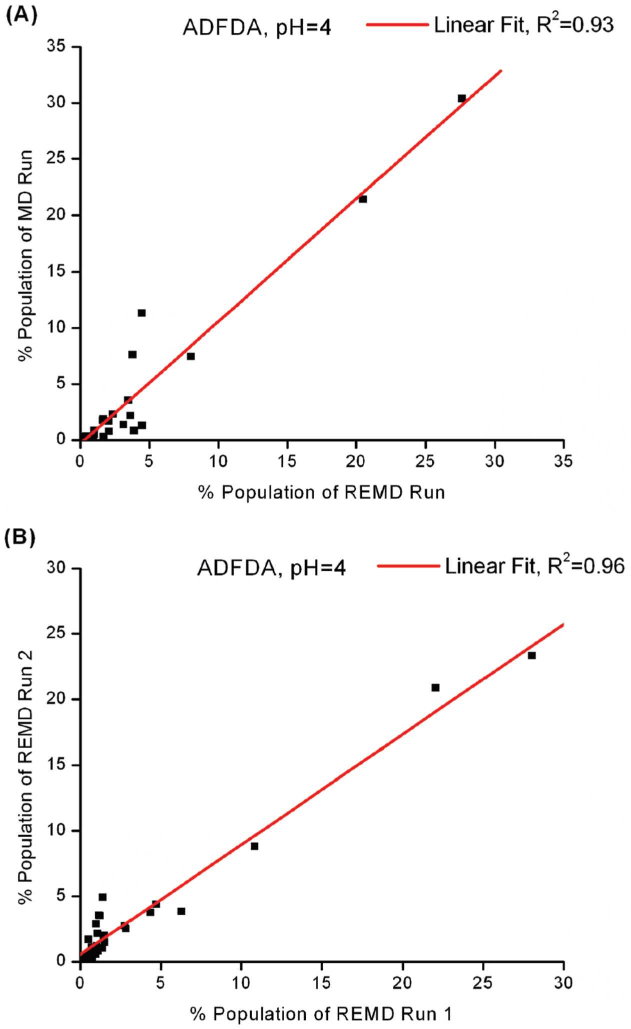


***Figure 6.*** Cluster populations of ADFDA at 300 K. (A) MD vs REMD at pH 4, (B) two REMD runs from different starting structures at pH 4. Large correlation shown in Figure 6B suggests that the REMD runs are converged. Large correlations between two independent REMD runs are also observed at other solution pH values. Correlations between MD and REMD simulations can be found in Table. 3.

the same conformational space and generated the same structure ensemble. The small size of ADFDA and simple structure of each residue make 100 ns long enough for MD to sample the relevant conformations.

We further studied the convergence of REMD simulations by comparing global conformation distribution between two REMD simulations starting from two different structures. Cluster populations of the two REMD simulations at solution pH 4 are displayed in Figure 6B. The $R^2$ value is 0.959 at pH 4. This large correlation tells us that the two REMD simulations provide the same structure ensemble, and hence the two simulations are converged.

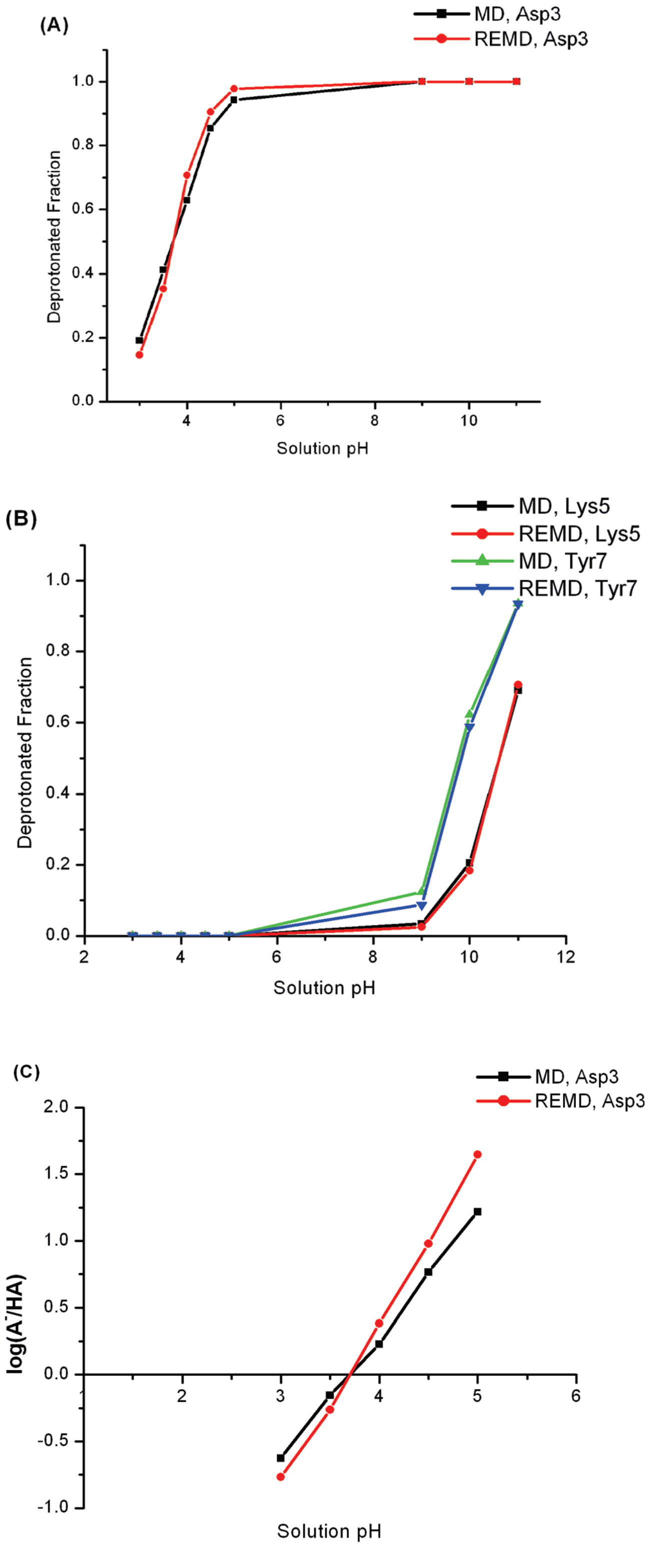


***Figure 7.*** (A and B) Titration curves of Asp3, Lys5, and Tyr7 in the heptapeptide derived from protein OMTKY3. (C) Hill's plots of Asp3. The p$K_a$ values of Asp3 are found through Hill's plots.

**C. Heptapeptide Derived from OMTKY3.** We first compared the protonation state sampling between constant pH REMD and MD simulations. Titration curves of Asp3, Lys5, and Tyr7 from two sets of simulations are plotted in Figure 7A and B. For each titratable residue, titration curves generated by constant pH REMD and MD are close to each other. Since the p$K_a$ value of Asp3 in this heptapeptide is experimentally determined to be 3.6, it will be interesting to evaluate how our predicted values compare to the experimental result. The p$K_a$ values of Asp3 were calculated on the basis of Hill's plots, which are displayed in Figure 7C. The predicted p$K_a$ value is 3.7 for both REMD and MD simulations, and they are in excellent agreement with the

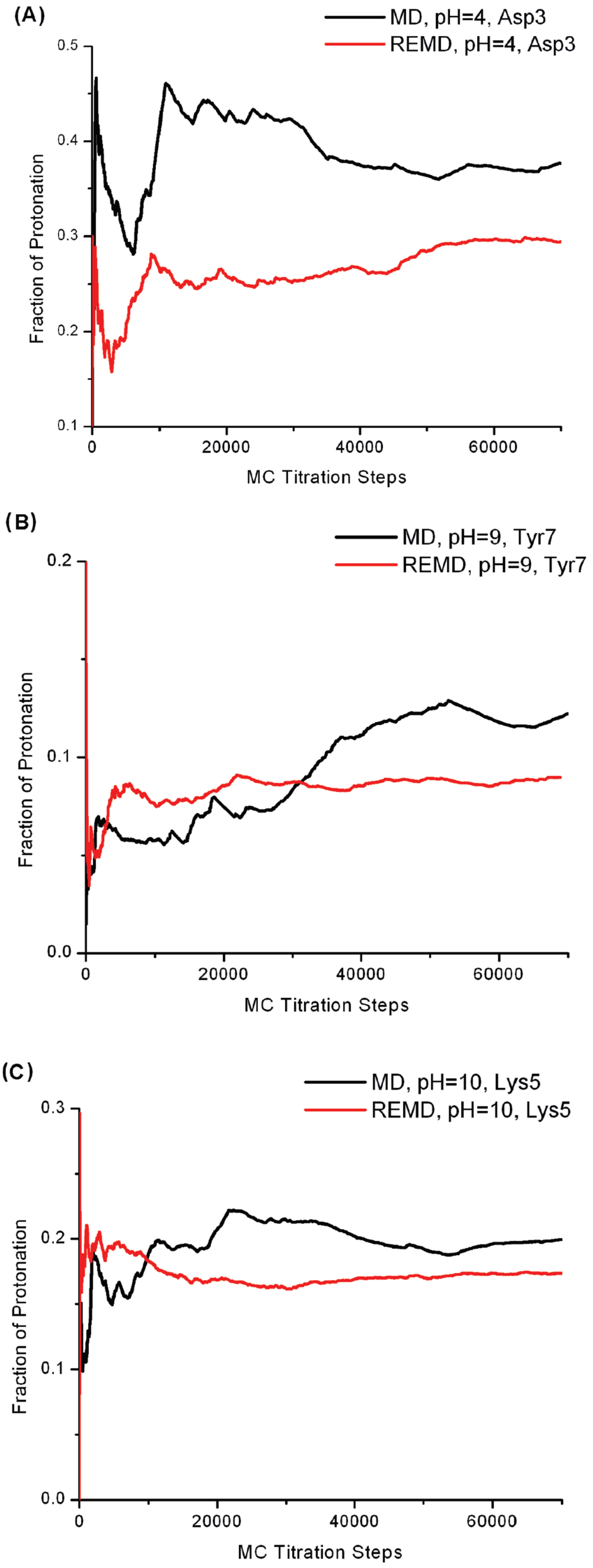


***Figure 8.*** Cumulative average protonation fraction of a titratable residue vs MC steps.

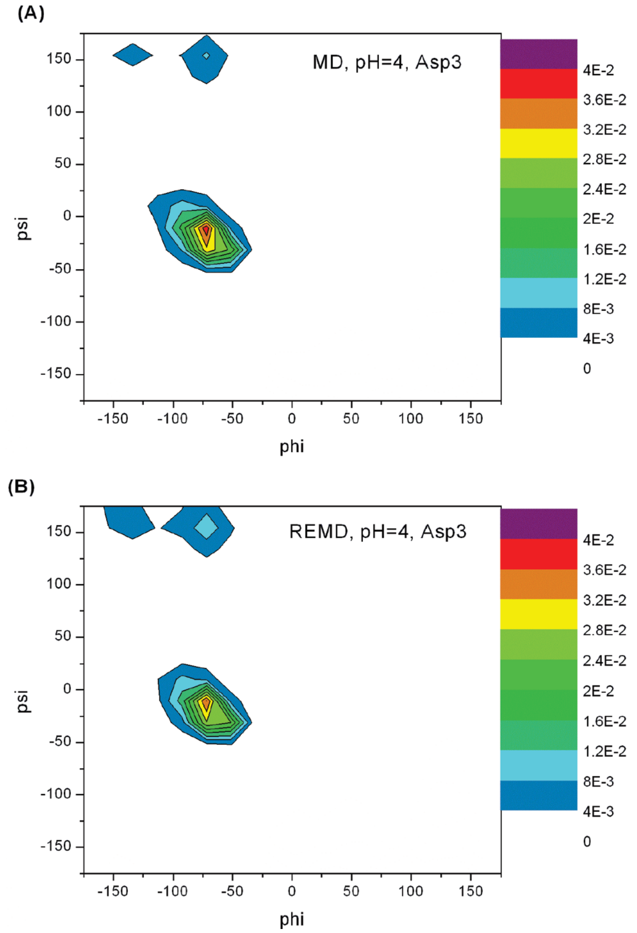


***Figure 9.*** Dihedral angle ($\varphi$, $\Psi$) probability densities of Asp3 at pH 4. (A) Constant-pH MD results; (B) constant-pH REMD results. All others also show a very similar trend.

experimental p$K_a$ value. Following the same procedures, our predicted p$K_a$ values of Lys5 (10.6 for both REMD and MD) and Tyr7 (9.9 and 9.8 for REMD and MD respectively) were obtained. Not surprisingly, the REMD and MD schemes yielded essentially the same predicted p$K_a$ values for Lys5 and Tyr7.

Although the final p$K_a$ predictions are the same for constant pH REMD and MD simulations, constant pH REMD showed a clear advantage in the convergence of protonation state sampling. Again, we chose the cumulative average protonation fraction vs MC steps to reflect protonation state sampling convergence for all three titratable residues. Several representative plots are shown in Figure 8. The trend that constant pH REMD simulations produce faster convergence in protonation fraction is universal. Therefore, it is very clear that constant pH REMD method is better than constant pH MD in protonation state sampling.

Conformational sampling is an important issue in constant pH studies. We first looked at the conformational sampling on peptide backbones. We evaluated backbone conformational sampling through Ramachandran plots. Six residues (from Ser2 to Tyr7) are studied here. Not surprisingly, Ramachandran plots from constant pH REMD and MD simulations are very close, suggesting that the overall local conformational samplings are similar. The Ramachandran plots of Asp3 at pH 4 are shown in Figure 9 as examples. The only exception is Tyr7 in acidic pH values. Tyr7 can visit the left-handed α-helix conformation during constant pH REMD runs but is not able to do that in constant pH

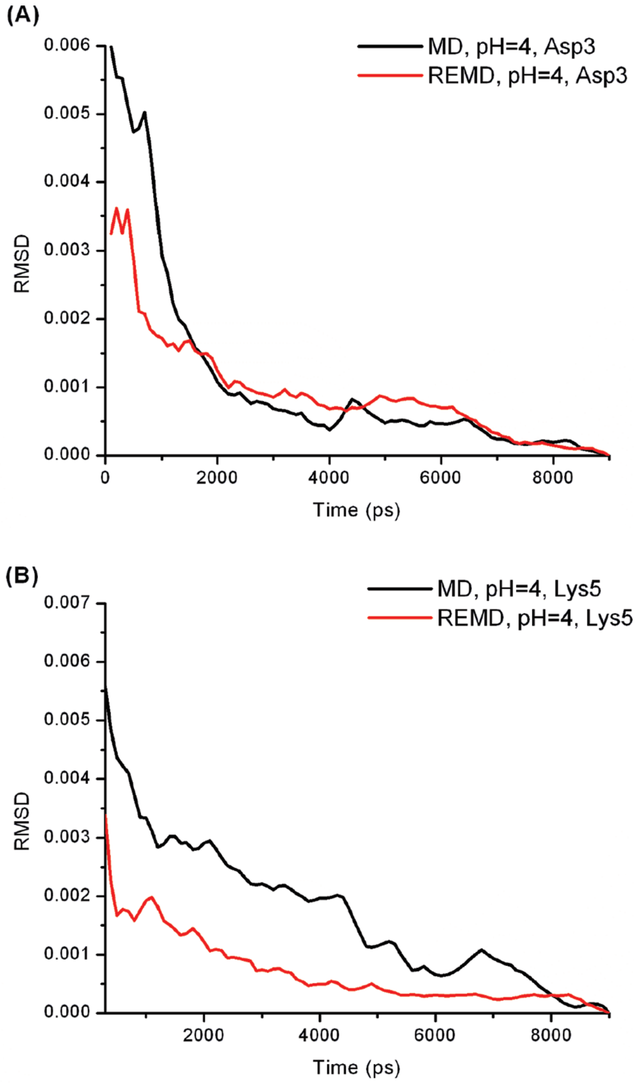


***Figure 10.*** The root-mean-square deviations (RMSDs) between the cumulative ($\varphi$, $\Psi$) probability density up to current time and the ($\varphi$, $\Psi$) probability density produced by the entire simulation. ($\varphi$, $\Psi$) probability density convergence behaviors at other pH values also show that REMD runs converge to final distribution faster.

MD runs. In general, constant pH REMD and MD yielded the same Ramachandran plots for the heptapeptide.

As demonstrated earlier, the overall samplings of ($\varphi$, $\Psi$) distribution by constant pH REMD and MD are similar for Ser2 to Thr6. It is interesting to determine how fast each sampling scheme reaches the final distribution. We studied the evolution of backbone conformational sampling based on cumulative data as we did in the case of protonation state sampling convergence. As described in the Methods subsection F, the RMSD between the ($\varphi$, $\Psi$) probability distribution up to current time versus total simulation time was calculated. The smaller a RMSD is, the closer a probability distribution reaches to the final distribution. Deviations were calculated starting from the second nanosecond with time intervals incremented by 100 ps. The cumulative time-dependence RMSD of Asp3 and Lys5 are also shown in Figure 10 as examples. As seen in the figures, these curves decrease faster in constant pH REMD simulations. Figure 10 suggests that, although the final ($\varphi$, $\Psi$) probability distributions are similar between constant pH REMD and MD simulations, the constant pH REMD simulation clearly reaches the final state faster.

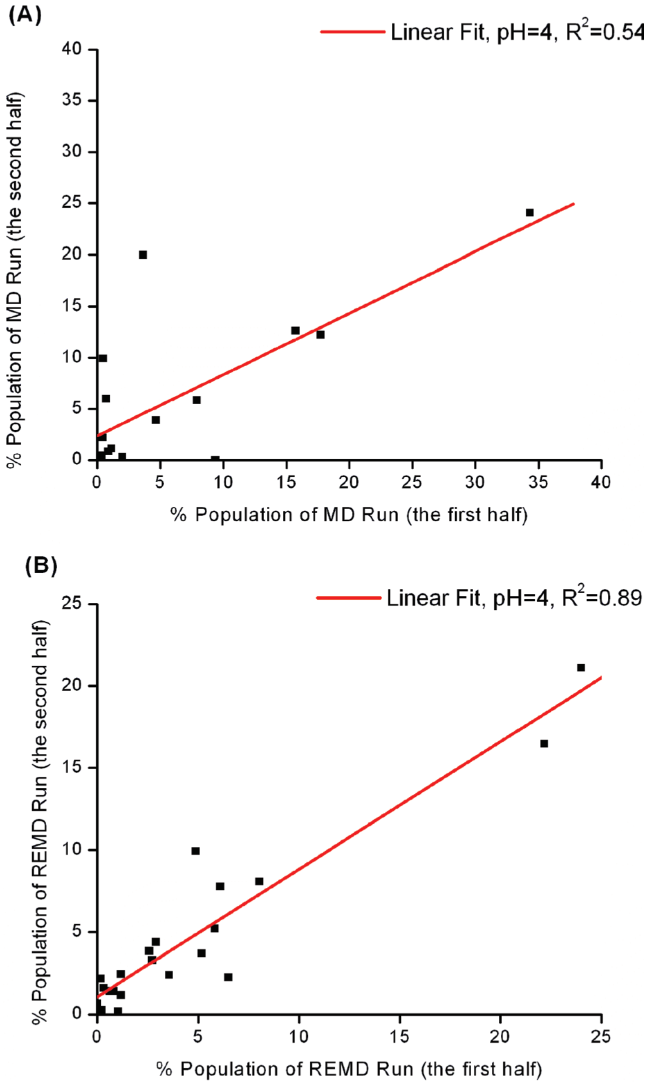


***Figure 11.*** Cluster population at 300 K from constant pH MD and REMD simulations at pH = 4. Cluster analysis is performed using the entire simulation. The populations in each cluster from the first and second halves of the trajectory are compared and plotted. Ideally, a converged trajectory should yield a correlation coefficient to be 1. (A) Constant pH MD. (B) Constant pH REMD. A much higher correlation coefficient can be seen in constant pH REMD simulation, suggesting much better convergence is achieved by the constant pH REMD run.

Cluster analysis was also applied to study the convergence of conformation sampling in the heptapeptide. By comparing cluster populations between the first and second half of one trajectory, one could check the convergence of that simulation. The two halves of a structural ensemble should yield the same populations in each cluster if convergence is reached. For example, for simulations at pH 4, both constant pH REMD and MD yield about 20 clusters, and the correlations coefficients are calculated through a linear regression. Cluster population plots and correlation coefficients are shown in Figure 11. A much higher correlation coefficient can be seen in constant pH REMD simulation, suggesting the two halves of the constant pH REMD simulation at pH 4 populate each cluster much more similarly than the corresponding constant pH MD does. Hence, much better convergence is achieved by the constant pH REMD run.

## Conclusion

In our work, we have applied the replica exchange molecular dynamics (REMD) algorithm to the discrete protonation state model developed by Mongan et al.[62] in order to study pH-dependent protein structure and dynamics. Seven small peptides were selected to test our constant pH REMD method. Constant pH molecular dynamics (MD) simulations were run on the same peptides for comparison. The constant REMD method results are encouraging. The constant REMD method can predict p$K_a$ values in agreement with literature and experimental results. The constant pH REMD method also displays an advantage in convergence behaviors during protonation states and conformational sampling.

The REMD algorithm has been proven beneficial to study pH-dependent protein structures. Our future work will include studies of pH-dependent protein dynamics and application of this constant pH REMD to large proteins.

**Acknowledgment.** We thank Carlos Simmerling and Wei Yang for helpful discussions. This work is supported from National Institute of Health under Contract 1R01 AI073674. Computer resources and support were provided by the Large Allocations Resource Committee through grant TG-MCA05S010 and the University of Florida High-Performance Computing Center.

### References

(1) Matthew, J. B.; Gurd, F. R. N.; Garciamoreno, E. B.; Flanagan, M. A.; March, K. L.; Shire, S. J. *Crc Cr. Rev. Bioch. Mol.* **1985**, *18*, 91–197.

(2) Mongan, J.; Case, D. A. *Curr. Opin. Struct. Biol.* **2005**, *15*, 157–163.

(3) Yang, A. S.; Honig, B. *J. Mol. Biol.* **1993**, *231*, 459–474.

(4) Bierzynski, A.; Kim, P. S.; Baldwin, R. L. *Proc. Natl. Acad. Sci. U. S. A.* **1982**, *79*, 2470–2474.

(5) Shoemaker, K. R.; Kim, P. S.; Brems, D. N.; Marqusee, S.; York, E. J.; Chaiken, I. M.; Stewart, J. M.; Baldwin, R. L. *Proc. Natl. Acad. Sci. U. S. A.* **1985**, *82*, 2349–2353.

(6) Schaefer, M.; Van Vlijmen, H. W. T.; Karplus, M. *Adv. Protein Chem.* **1998**, *51*, 1–57.

(7) Antosiewicz, J.; Briggs, J. M.; McCammon, J. A. *Eur. Biophys. J. Biophy.* **1996**, *24*, 137–141.

(8) Hunenberger, P. H.; Helms, V.; Narayana, N.; Taylor, S. S.; McCammon, J. A. *Biochemistry* **1999**, *38*, 2358–2366.

(9) Demchuk, E.; Genick, U. K.; Woo, T. T.; Getzoff, E. D.; Bashford, D. *Biochemistry* **2000**, *39*, 1100–1113.

(10) Dillet, V.; Dyson, H. J.; Bashford, D. *Biochemistry* **1998**, *37*, 10298–10306.

(11) Harris, T. K.; Turner, G. J. *IUBMB Life* **2002**, *53*, 85–98.

(12) Kelly, J. W. *Curr. Opin. Struct. Biol.* **1996**, *6*, 11–17.

(13) Kelly, J. W. *Structure* **1997**, *5*, 595–600.

(14) Rastogi, V. K.; Girvin, M. E. *Nature* **1999**, *402*, 263–268.

(15) Hill, T. L. *J. Am. Chem. Soc.* **1956**, *78*, 3330–3336.

(16) Simonson, T.; Carlsson, J.; Case, D. A. *J. Am. Chem. Soc.* **2004**, *126*, 4167–4180.

(17) Tanford, C.; Kirkwood, J. G. *J. Am. Chem. Soc.* **1957**, *79*, 5333–5339.

(18) Warshel, A. *Nature* **1987**, *330*, 15–16.

(19) Langsetmo, K.; Fuchs, J. A.; Woodward, C. *Biochemistry* **1991**, *30*, 7603–7609.

(20) Eberini, I.; Baptista, A. M.; Gianazza, E.; Fraternali, F.; Beringhelli, T. *Proteins: Struct., Funct., Bioinf.* **2004**, *54*, 744–758.

(21) Sham, Y. Y.; Muegge, I.; Warshel, A. *Biophys. J.* **1998**, *74*, 1744–1753.

(22) Simonson, T.; Archontis, G.; Karplus, M. *J. Phys. Chem. B* **1999**, *103*, 6142–6156.

(23) Warshel, A.; Aqvist, J. *Annu. Rev. Biophys. Biomol. Struct.* **1991**, *20*, 267–298.

(24) Antosiewicz, J.; Mccammon, J. A.; Gilson, M. K. *J. Mol. Biol.* **1994**, *238*, 415–436.

(25) Antosiewicz, J.; McCammon, J. A.; Gilson, M. K. *Biochemistry* **1996**, *35*, 7819–7833.

(26) Bashford, D.; Karplus, M. *Biochemistry* **1990**, *29*, 10219–10225.

(27) Demchuk, E.; Wade, R. C. *J. Phys. Chem.* **1996**, *100*, 17373–17387.

(28) Kamerlin, S. C. L.; Haranczyk, M.; Warshel, A. *J. Phys. Chem. B* **2009**, *113*, 1253–1272.

(29) Riccardi, D.; Schaefer, P.; Cui, Q. *J. Phys. Chem. B* **2005**, *109*, 17715–17733.

(30) Warshel, A.; Sussman, F.; Hwang, J. K. *J. Mol. Biol.* **1988**, *201*, 139–159.

(31) Bas, D. C.; Rogers, D. M.; Jensen, J. H. *Proteins: Struct., Funct., Bioinf.* **2008**, *73*, 765–783.

(32) Li, H.; Robertson, A. D.; Jensen, J. H. *Proteins: Struct., Funct., Bioinf.* **2005**, *61*, 704–721.

(33) Alexov, E. G.; Gunner, M. R. *Biophys. J.* **1997**, *72*, 2075–2093.

(34) Baptista, A. M. *J. Chem. Phys.* **2002**, *116*, 7766–7768.

(35) Baptista, A. M.; Martel, P. J.; Petersen, S. B. *Proteins* **1997**, *27*, 523–544.

(36) Baptista, A. M.; Teixeira, V. H.; Soares, C. M. *J. Chem. Phys.* **2002**, *117*, 4184–4200.

(37) Georgescu, R. E.; Alexov, E. G.; Gunner, M. R. *Biophys. J.* **2002**, *83*, 1731–1748.

(38) Jensen, J. H.; Li, H.; Robertson, A. D.; Molina, P. A. *J. Phys. Chem. A* **2005**, *109*, 6634–6643.

(39) Khandogin, J.; Brooks, C. L. *Biophys. J.* **2005**, *89*, 141–157.

(40) Khandogin, J.; Brooks, C. L. *Biochemistry* **2006**, *45*, 9363–9373.

(41) Khandogin, J.; Brooks, C. L. *Proc. Natl. Acad. Sci. U. S. A.* **2007**, *104*, 16880–16885.

(42) Khandogin, J.; Chen, J. H.; Brooks, C. L. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 18546–18550.

(43) Khandogin, J.; Raleigh, D. P.; Brooks, C. L. *J. Am. Chem. Soc.* **2007**, *129*, 3056–3057.

(44) Lee, A. C.; Yu, J. Y.; Crippen, G. M. *J. Chem. Inf. Model.* **2008**, *48*, 2042–2053.

(45) Livesay, D. R.; Jacobs, D. J.; Kanjanapangka, J.; Chea, E.; Cortez, H.; Garcia, J.; Kidd, P.; Marquez, M. P.; Pande, S.; Yang, D. *J. Chem. Theory Comput.* **2006**, *2*, 927–938.

(46) Machuqueiro, M.; Baptista, A. M. *J. Phys. Chem. B* **2006**, *110*, 2927–2933.

(47) Machuqueiro, M.; Baptista, A. M. *Biophys. J.* **2007**, *92*, 1836–1845.

(48) Machuqueiro, M.; Baptista, A. M. *Proteins: Struct., Funct., Bioinf.* **2008**, *72*, 289–298.

(49) Machuqueiro, M.; Baptista, A. M. *J. Am. Chem. Soc.* **2009**, *131*, 12586–12594.

(50) Madura, J. D.; Briggs, J. M.; Wade, R. C.; Davis, M. E.; Luty, B. A.; Ilin, A.; Antosiewicz, J.; Gilson, M. K.; Bagheri, B.; Scott, L. R.; Mccammon, J. A. *Comput. Phys. Commun.* **1995**, *91*, 57–95.

(51) Minikis, R. M.; Kairys, V.; Jensen, J. H. *J. Phys. Chem. A* **2001**, *105*, 3829–3837.

(52) Mertz, J. E.; Pettitt, B. M. *Int. J. Supercomput. Ap.* **1994**, *8*, 47–53.

(53) Borjesson, U.; Hunenberger, P. H. *J. Chem. Phys.* **2001**, *114*, 9706–9719.

(54) Borjesson, U.; Hunenberger, P. H. *J. Phys. Chem. B* **2004**, *108*, 13551–13559.

(55) Lee, M. S.; Salsbury, F. R.; Brooks, C. L. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 738–752.

(56) Kong, X. J.; Brooks, C. L. *J. Chem. Phys.* **1996**, *105*, 2414–2423.

(57) Burgi, R.; Kollman, P. A.; van Gunsteren, W. F. *Proteins* **2002**, *47*, 469–480.

(58) Dlugosz, M.; Antosiewicz, J. M. *Chem. Phys.* **2004**, *302*, 161–170.

(59) Dlugosz, M.; Antosiewicz, J. M. *J. Phys. Chem. B* **2005**, *109*, 13777–13784.

(60) Dlugosz, M.; Antosiewicz, J. M. *J. Phys.: Condens. Matter* **2005**, *17*, S1607–S1616.

(61) Dlugosz, M.; Antosiewicz, J. M.; Robertson, A. D. *Phys. Rev. E* **2004**, *69*, 021915.

(62) Mongan, J.; Case, D. A.; McCammon, J. A. *J. Comput. Chem.* **2004**, *25*, 2038–2048.

(63) Walczak, A. M.; Antosiewicz, J. M. *Phys. Rev. E* **2002**, *66*, 051911.

(64) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087–1092.

(65) Bashford, D.; Case, D. A. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129–152.

(66) Onufriev, A.; Bashford, D.; Case, D. A. *J. Phys. Chem. B* **2000**, *104*, 3712–3720.

(67) Hamelberg, D.; Mongan, J.; McCammon, J. A. *J. Chem. Phys.* **2004**, *120*, 11919–11929.

(68) Hamelberg, D.; Mongan, J.; McCammon, J. A. *Protein Sci.* **2004**, *13*, 76–76.

(69) Williams, S. L.; de Oliveira, C. A. F.; McCammon, J. A. *J. Chem. Theory Comput.* **2010**, *6*, 560–568.

(70) Li, H. Z.; Fajer, M.; Yang, W. *J. Chem. Phys.* **2007**, *126*, 024106.

(71) Mitsutake, A.; Sugita, Y.; Okamoto, Y. *Biopolymers* **2001**, *60*, 96–123.

(72) Zheng, L. Q.; Chen, M. G.; Yang, W. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 20227–20232.

(73) Zheng, L. Q.; Chen, M. G.; Yang, W. *J. Chem. Phys.* **2009**, *130*, 234105.

(74) Berg, B. A.; Neuhaus, T. *Phys. Lett. B* **1991**, *267*, 249–253.

(75) Berg, B. A.; Neuhaus, T. *Phys. Rev. Lett.* **1992**, *68*, 9–12.

(76) Lyubartsev, A. P.; Martsinovski, A. A.; Shevkunov, S. V.; Vorontsovvelyaminov, P. N. *J. Chem. Phys.* **1992**, *96*, 1776–1783.

(77) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

(78) Hansmann, U. H. E. *Chem. Phys. Lett.* **1997**, *281*, 140–150.

(79) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668–1688.

(80) Bashford, D.; Case, D. A.; Dalvit, C.; Tennant, L.; Wright, P. E. *Biochemistry* **1993**, *32*, 8045–8056.

(81) Case, D. A.; Darden, T. A.; Cheatham, I.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R. C.; Zhang, W.; Merz, K. M.; B.Wang; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A. *AMBER 10*; University of California: San Francisco, 2008.

(82) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 712–725.

(83) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327–341.

(84) Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(85) Elber, R.; Roitberg, A.; Simmerling, C.; Goldstein, R.; Li, H. Y.; Verkhivker, G.; Keasar, C.; Zhang, J.; Ulitsky, A. *Comput. Phys. Commun.* **1995**, *91*, 159–189.

(86) Okur, A.; Roe, D. R.; Cui, G. L.; Hornak, V.; Simmerling, C. *J. Chem. Theory Comput.* **2007**, *3*, 557–568.

(87) Okur, A.; Wickstrom, L.; Layten, M.; Geney, R.; Song, K.; Hornak, V.; Simmerling, C. *J. Chem. Theory Comput.* **2006**, *2*, 420–433.

CT900676B

JCTC Journal of Chemical Theory and Computation

# Case Studies of ONIOM(DFT:DFTB) and ONIOM(DFT:DFTB:MM) for Enzymes and Enzyme Mimics

Marcus Lundberg,[†,§] Yoko Sasakura,[†] Guishan Zheng,[‡,||] and Keiji Morokuma*[,†,‡]

*Fukui Institute for Fundamental Chemistry, Kyoto University, 34-4 Takano Nishihiraki-cho, Sakyo, Kyoto 606-8103, Japan, and Cherry L. Emerson Center for Scientific Computation and Department of Chemistry, Emory University, Atlanta, Georgia 30322*

Received January 14, 2010

**Abstract:** The replacement of standard molecular mechanics force fields by inexpensive molecular orbital (QM′) methods in multiscale models has many advantages, e.g., a more straightforward description of mutual polarization and charge transfer between layers. The ONIOM(QM:QM′) scheme with mechanical embedding can combine any two methods without prior parametrization or significant coding effort. In this scheme, the environmental effect is evaluated fully at the QM′ level, and the accuracy therefore depends on how well the low-level QM′ method describes the changes in electron density of the reacting region. To examine the applicability of the QM:QM′ approach, we perform case studies with density-functional tight-binding (DFTB) as the low-level QM′ method in two-layer ONIOM(B3LYP/6-31G(d):DFTB) models. The investigated systems include simple amino acid models, one nonheme iron enzyme mimic, and the enzymatic reactions of Zn-$\beta$-lactamase and trypsin. For the last example, we also illustrate the use of a three-layer ONIOM(B3LYP/6-31G(d):DFTB:Amber96) model. The ONIOM extension, compared to the QM calculation for the small model system, improves the relative energies, but high accuracy (deviations below 1 kcal/mol) is not achieved even with relatively large QM models. Polarization effects are fairly well described using DFTB, but in some cases QM and QM′ methods converge to different electronic states. We discuss when the QM:QM′ approach is appropriate and the possibilities of estimating the quality of the ONIOM extension without having to make explicit benchmarks of the entire system.

## I. Introduction

Over the three decades following their first implementation,[1] the quantum mechanics/molecular mechanics (QM/MM) methods have seen great success in a wide spectrum of applications, including biological reactions and materials science.[2−5] The success of the QM/MM methods is rooted in their multiscale nature, in which the system is partitioned into different regions treated at appropriate levels of theory. Only the core region, i.e., the QM region, is treated with a computationally expensive QM method that can describe chemical reactions, e.g., *ab initio* wave function, density functional, or semiempirical methods, while the rest of the system, i.e., the MM region, is treated with MM methods that are often thousands of times faster than QM methods. This compromise between accuracy and computational efficiency makes it possible to study systems that are computationally prohibitive to the pure QM methods.

The success of QM/MM has stimulated our development of the multiscale ONIOM method.[6−12] In two-layer ONIOM, the *high*-level method (QM) is applied only to a selected *model* system, i.e., the QM region with link atoms.[10,13] The

* Corresponding author telephone: +81-75-711-7843; fax: +81-75-781-4757; e-mail: morokuma@fukui.kyoto-u.ac.jp.

[†] Kyoto University.

[‡] Emory University.

[§] Present address: Department of Chemistry, Stanford University, Stanford, California 94305.

[||] Present address: Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138.

10.1021/ct100029p © 2010 American Chemical Society
Published on Web 03/16/2010

*low*-level method is applied to both the *model* and the *real* system. The latter is equivalent to the entire system and includes both the QM and MM regions of generic QM/MM models. The energy of the target *real*,*high* calculation is then approximated by an extrapolation scheme. An advantage of the ONIOM scheme is that all calculations are made on complete systems, which makes it possible to combine any methods, including molecular orbital methods, without prior parametrization or significant coding efforts. The method can also easily be extended to an arbitrary number of layers, although the present implementation is limited to three layers.[9,14−16] The ONIOM method has been successfully utilized in a wide range of applications, including transition metal catalysis, carbon nanotube chemistry, and enzymatic catalysis.[17,18]

Although the ONIOM method is different in design and implementation compared to generic QM/MM methods, there are also many similarities. When comparing these methods, we use the general terms QM region and MM region to represent different partitions of the system. One common concern is the boundary between QM and MM regions, and how to treat the interactions between the two. Some issues arising from the use of two fundamentally different physical descriptions of the system can be listed as follows: (1) the exchange interaction between the QM region and the MM region, which is partially included in the parametrized van der Waals interaction between the two regions, (2) the charge transfer between the QM and the MM region, (3) the polarization of the QM electron density induced by the MM atoms, and (4) the polarization of the MM atoms by QM and other MM atoms. Among these four issues, the last one is often considered the most severe and important.[19−25]

To properly address the polarization of MM atoms, the most straightforward approach (but by no means trivial) is to develop a polarizable force field.[19,20,26] In recent years, several approaches have been adopted to develop such force fields, e.g., the fluctuation charge model[19,20] and explicit polarization potential.[27] These models can partially alleviate the issues connected with the neglect of polarization effects, but they are not without problems.[19,20] For example, dissociated diatomic molecules bear finite charges on both atoms. This has been explicitly addressed by the Martinez group by using the QTPIE model, a generalization of the flucq model.[21,23,24] Although there is much progress, the performance of currently proposed polarizable force field models remains to be seen.

Development of accurate force fields for biological systems is further complicated by the large number of parameters.[22,28] QM/MM calculations have also been extended to include charge transfer between QM and MM regions using the principle of chemical potential equilibration.[29] A drawback of these methods is that they require multiple QM iterations to reach consistency between QM and MM charges, which leads to significant increases in computational cost.

An alternative approach is to describe the entire system by quantum mechanics. Large systems can be treated entirely by high-level QM methods by dividing them into fragments,[30−32] but these methods invariably spend the largest amount of time calculating the nonreacting part of the system. Savings in computational time can be achieved by a QM/QM′ approach, where a less expensive molecular orbital method, QM′, is used to describe the environment. In early work, Cortona proposed to combine different density functionals based on superposition of atomic densities.[33] Other groups have used the QM/QM′ approach to embed an expensive correlated wave function calculation in the environment of a relatively fast DFT method.[34−36] To be able to treat very large systems, Gogonea et al. constructed a hybrid DFT/semiempirical Hamiltonian based on the idea of equilibration of chemical potentials combined with the divide and conquer method.[30,37] The approach allows for mutual polarization and charge transfer but requires an iterative approach to equilibrate the chemical potential. Cui et al. have also coupled DFT with semiempirical methods, both in an iterative fashion and in the ONIOM formalism.[38]

The properties of the ONIOM method make it very straightforward to design QM:QM′ models.[7,10,11,39] As no parametrization is required to combine different methods, the selection of quantum mechanical methods can be made to fit each specific chemical process.[14,16,40,41] We are interested in reactions in complex systems, e.g., transition metal enzymes. For these systems, the compromise is often to use the best possible method to calculate the reaction energy and use a relatively cheap method to describe the environmental effects. In the present study, QM is a density functional method, while QM′ is the density functional tight binding method (DFTB). In the ONIOM(QM:QM′) scheme with mechanical embedding (ME), the *model*,*QM* calculations are independent of the QM′ electron density, and there is no need for multiple iterations of the QM region. This makes the QM:QM′−ME approach cost-efficient for systems that require expensive QM schemes, as it minimizes the number of QM iterations.

A limitation in the mechanical embedding scheme is that the environmental effects are evaluated only at the QM′ (*low*) level, as the difference between *real*,*QM′* and *model*,*QM′* electron densities. The QM′ method must therefore be able to describe the changes in electron density of the reacting region; e.g., if an electron transfer reaction occurs in the *model*,*QM* system, the same process should occur also in the layers described by QM′. The applicability of the QM:QM′ scheme thus depends heavily on the QM′ description of the environmental effect on a specific reaction, and it is difficult to perform comprehensive benchmark tests. The purpose of this investigation is therefore to demonstrate the advantages and limitations of the QM:QM′ method by applying it to a few illustrative examples. The resultant understanding of the principles of the ONIOM scheme should help in the design of more reliable QM:QM′ methods for a wide range of systems than before.

Another important aspect of the QM:QM′ scheme is that QM′ calculation is performed for the entire *real* system, which means that parametrized methods used as QM′ must have parameters applicable also for the reactive region. In the present paper, for reactions including transition metals, we propose to use a density functional method as the high-level QM method and the density-functional tight-binding (DFTB) method[42−44] as the low-level QM′ method. The

DFTB method is a second-order approximation of DFT;[45] this combination can be expected to give a similar performance to pure DFT methods.[46,47] Since DFTB is 100−1000 times faster than DFT, the ONIOM combination can be applied to a much larger system than a pure DFT method can be applied to. We have developed DFTB parameters for many of the first-row transition metal elements[48] and have also implemented a DFTB code in the Gaussian program, which enables us to use the full functionalities of this program for DFTB calculations.[47]

We have previously illustrated the applicability of the DFT:DFTB approach for transition metal systems by applying it to a nonheme enzyme (isopenicillin N synthase).[48] DFTB has also been used as a *low* layer in B3LYP:DFTB and MP2:DFTB combinations for copper- and titanium-containing systems, with errors typically less than 2 kcal/mol.[49] The choice of DFTB as the QM′ method for proteins is supported by the good performance for geometries and relative energies in biological systems.[50−52]

The QM:QM′−ME approach is efficient only when the cost of the QM calculation for the *model* system is higher than the cost of the QM′ calculation for the *real* system. For very large systems, it is therefore useful to make a three-layer partition of the system, i.e., QM:QM′:MM.[9] As a computationally cheap QM′ method can be utilized economically with thousands of atoms in many applications, this method pushes the boundary between QM and MM away from the reaction center, which is expected to largely reduce the issues caused by the QM:MM boundary mentioned above. The performance of three-layer ONIOM has been tested by our group before.[14−16] Other three-layer approaches have been used to describe, e.g., spectral tuning in protein environments.[53,54]

In this study, we test both two-layer ONIOM(B3LYP:DFTB) and three-layer ONIOM(B3LYP:DFTB:MM) models. In section II, we discuss how to evaluate ONIOM in terms of its performance compared to the *high*-level calculation of the *real* system. In the following section, several carefully chosen examples are presented, which range from proton affinity calculations of titratable amino acids to an active site model of Zn-$\beta$-lactamase, a nonheme iron catalase mimic, and finally to the acylation process in trypsin. A discussion of the advantages and limitations of the ONIOM(QM:QM′) scheme is presented in section IV before the conclusions in section V.

## II. Computational Details

**A. ONIOM Calculation Evaluation.** ONIOM uses an extrapolation scheme to approximate the costly target calculation, a *high* (QM)-level treatment of the entire *real* system:

$$\Delta E^{\mathrm{ONIOM}} = \Delta E^{\mathrm{model,high}} + \Delta E^{\mathrm{real,low}} - \Delta E^{\mathrm{model,low}} \tag{1}$$

where $\Delta E$ refers to the energy of reaction, $E$(product) − $E$(reactant), or the activation energy, $E$(transition state) − $E$(reactant).

The error $\Delta D$ in $\Delta E$ in the ONIOM approximation, relative to the *real*,*high* target calculation, can be written as

$$\Delta D = \Delta E^{\mathrm{real,high}} - \Delta E^{\mathrm{ONIOM}} \tag{2}$$

A convenient measure for evaluating ONIOM calculations is the $\Delta S$ value that describes the environmental effect at each computational level:[11]

$$\Delta S^{\mathrm{high}} = \Delta E^{\mathrm{real,high}} - \Delta E^{\mathrm{model,high}} \tag{3}$$

$$\Delta S^{\mathrm{low}} = \Delta E^{\mathrm{real,low}} - \Delta E^{\mathrm{model,low}} \tag{4}$$

$\Delta S^{\mathrm{high}}$ is "the substituent effect" on the energetics when the *real* system is approximated by the *model* system at the given *high*-level method, i.e., a measure of environmental effect; $\Delta S^{\mathrm{low}}$ is the corresponding value for the *low*-level method. Using eqs 1−4, the error in the ONIOM approximation can be conveniently expressed in the form of $\Delta S$ values:

$$\begin{aligned}\Delta D &= \Delta E^{\mathrm{real,high}} - \Delta E^{\mathrm{ONIOM}}\\ &= \Delta E^{\mathrm{real,high}} - (\Delta E^{\mathrm{model,high}} + \Delta E^{\mathrm{real,low}} - \Delta E^{\mathrm{model,low}})\\ &= \Delta S^{\mathrm{high}} - \Delta S^{\mathrm{low}}\end{aligned} \tag{5}$$

If the low-level method describes the environmental effect in the same way as the high-level method ($\Delta S^{\mathrm{low}} = \Delta S^{\mathrm{high}}$), the ONIOM energy is exact, i.e., becomes the same as the target calculation, $\Delta E^{\mathrm{real,high}}$, with $\Delta D = 0$. The error can also be small if the total effect of the environment is very small ($\Delta S \approx 0$) at each level; in this case, a *high*-level calculation of the *model* system is enough, and there is no need for an ONIOM extension. To distinguish the situation where the error is small due to small environmental effects from the situation where the ONIOM extrapolation scheme is accurate, we introduce the ONIOM error score (OES):

$$\mathrm{OES} = \Delta D/\Delta S^{\mathrm{high}} \tag{6}$$

A small absolute value of the error score means that the ONIOM setup is appropriate. If the absolute value of the error score is <1, the ONIOM calculation improves the result compared to the *model* calculation. If the absolute value is >1, the addition of the *real* system makes the relative energies even worse. It is possible that the *model*,*high* calculation gives the right result by neglecting two different environmental effects with opposite sign at the two different levels. In these cases, an ONIOM model could show a larger error but still give a qualitatively more correct description of the system. The ONIOM error score is therefore mainly of use when evaluating isolated or incremental effects.

**B. Computational Details.** Calculations have been performed using a development version of the Gaussian 03 package[55] in which the DFTB method has been implemented.[47] Tests of the accuracy of the ONIOM model are made against the target *real*,*high* calculation for all systems, and to simplify the analysis, all comparisons between methods are made at the same geometry. A previous QM:DFTB test by Iordanov showed that geometry errors were in general smaller than energetic errors.[49]

In most calculations, the high-level method is B3LYP/6-31G(d), mainly because DFTB is parametrized against DFT with a double-$\zeta$ basis set. The contribution from the DFTB layer only reflects the environmental effect, and assuming

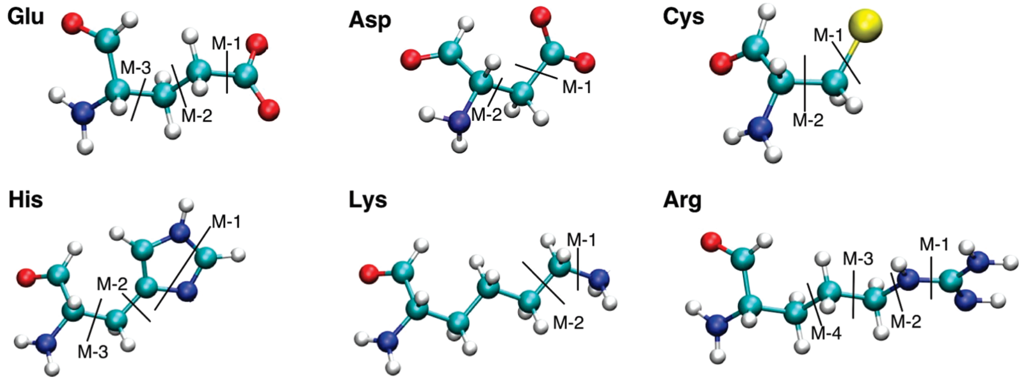


***Figure 1.*** ONIOM(QM:QM′) partitions in amino acid side chains. Lines labeled M-1 to M-4 illustrate model cuts, with the part to the right being part of the *model* system.

the selected high-level method gives a good description of that effect, the ONIOM error should not be very sensitive to the choice of high-level methods. In cases where the environmental effects are basis-set-dependent, e.g., in the calculation of proton affinities, calculations have also been made with larger basis sets.

## III. Results

The selected systems represent a wide range of applications, from simple amino acid models to medium-sized models of transition metal systems and a full protein. The purpose of calculating proton affinities for amino acids is to illustrate the general principles of the QM:QM′ approach for reactions with clear environmental effects. Tests of medium-sized models include transition metal systems, one enzymatic reaction, Zn-$\beta$-lactamase, and an iron catalase mimic. The latter example shows how the presence of multiple electronic states affects the applicability of the QM:QM′ scheme. Finally, we apply the three-layer QM:QM′:MM model to trypsin to illustrate that the method can be used for full enzymatic systems.

**A. Proton Affinities of Amino Acids and Peptides—Effect of Improved Treatment of Side Chains.** In active-site QM models of enzymatic reactions, amino acid residues are often represented by truncated models of their side chains; e.g., acetate is used as a model for the negative residue Asp.[56] Here, we evaluate the benefit of an additional QM′ layer to represent the left-out part of the side chain when calculating the gas phase proton affinity of six titratable side chains: arginine, lysine, histidine, aspartate, glutamate, and cysteine. Proton affinities are very sensitive to environmental effects and good candidates for evaluating different methods.[57]

*Amino Acid Monomers.* In the first set of calculations, the *real* systems are the side chains, with truncated backbone bonds capped by hydrogen atoms. Several *model* systems are designed by making different cuts in each side chain, see Figure 1. Proton affinities with pure DFTB have been calculated using an energy of $H^+$ of 141.9 kcal/mol.[58] ONIOM results are not affected by this value as it cancels in the ($E^{\text{real,low}} - E^{\text{model,low}}$) contribution. Results do not include corrections for basis set superposition error (BSSE). The $\Delta S^{\text{high}}$ values are not likely to be much affected by BSSE, as the substituents are made in sites away from the proton. Geometries are optimized using ONIOM separately for each size of the *model* system, and using B3LYP/6-31G(d) for the reference calculation. The energies from the respective *real*,*high* calculations for each ONIOM model can therefore be used to assess the quality of the ONIOM geometry optimization.

Results are summarized in Table 1. The ONIOM results are significant improvements compared to the *model*,*high* calculations (using the *real*,*high* calculations as a reference). As an example, the error for an acetate model (M-2) of glutamate is 10.8 kcal/mol at the *model*,*B3LYP* level and decreases to 2.2 kcal/mol with the additional DFTB layer (ONIOM error score of 0.21). For residues where the base is negative, i.e., Asp, Glu, and Cys, the ONIOM errors do not consistently decrease as the size of the *model* system increases. The remaining ONIOM error for the largest *model* systems comes from an underestimation of the effect of the backbone of 2−4 kcal/mol in DFTB compared to DFT. In the smallest model system, this error is partly canceled by a difference between DFTB and DFT in describing the effect of adding the first methyl group. This error cancellation gives a smaller apparent ONIOM error for the smallest *model* systems compared to the larger *model* systems.

In contrast, residues whose base is neutral, i.e., His, Lys, and Arg, show lower errors when the size of the *model* system increases, and errors for cuts three or more bonds away from the proton are ~1 kcal/mol or less. ONIOM optimization also gives overall good performance for geometries with typical deviations in *real*,*high* energy between different geometries of <1 kcal/mol. However, the artificially truncated backbone is not restricted and can adapt different conformations, which may lead to jumps in the calculated proton affinity. This explains the differences in the calculated proton affinity of the *real*,*B3LYP* system between Lys model M-2 and the full model, as well as differences between Arg model M-4 and the full model, see Table 1.

*Tripeptides.* To better include the effect of the peptide backbone, two tests were made for tripeptides. Histidine was chosen as a representative of the "neutral base" group, and glutamate was chosen as a representative of the "negative base" group. For the Gly−Glu−Gly (GEG) tripepeptide, the error when using acetate as a model (M-2) is 28.2 kcal/mol at the B3LYP level, and adding the DFTB layer reduces the error to 0.7 kcal/mol. Results are similar also for other cut positions, see Table 1. For the Gly−His−Gly (GHG)

***Table 1.*** Proton Affinities (kcal/mol) of Selected Amino Acids Calculated Using ONIOM B3LYP/6-31G(d):DFTB[a]

| real system | model system | B3LYP (real) | ONIOM | B3LYP (model) | DFTB (real) | DFTB (model) | $\Delta S^{\rm high}$ | $\Delta D$ | OES |
|---|---|---|---|---|---|---|---|---|---|
| Glu | M-1 | 355.6 | 355.2 | 362.5 | 358.4 | 365.7 | −6.9 | 0.4 | −0.05 |
| | M-2 | 355.6 | 357.9 | 366.5 | 358.0 | 366.6 | −10.8 | −2.2 | 0.21 |
| | M-3 | 355.3 | 357.5 | 364.8 | 358.4 | 365.7 | −9.5 | −2.2 | 0.23 |
| | full | **354.7** | | | 358.0 | | | | |
| Asp | M-1 | 352.3 | 352.8 | 362.5 | 356.0 | 365.7 | −10.2 | −0.5 | 0.05 |
| | M-2 | 352.1 | 355.5 | 366.3 | 355.8 | 366.6 | −14.1 | −3.3 | 0.24 |
| | full | **352.0** | | | 355.7 | | | | |
| Cys | M-1 | 353.2 | 350.4 | 363.0 | 343.9 | 356.5 | −9.7 | 2.9 | −0.29 |
| | M-2 | 353.4 | 357.1 | 369.1 | 344.1 | 356.1 | −15.7 | −3.7 | 0.23 |
| | full | **352.9** | | | 344.3 | | | | |
| His | M-1 | 247.2 | 252.6 | 218.3 | 239.5 | 205.2 | 28.9 | −5.5 | −0.19 |
| | M-2 | 246.2 | 244.4 | 236.6 | 239.5 | 231.7 | 9.6 | 1.9 | 0.19 |
| | M-3 | 246.4 | 245.2 | 240.6 | 239.5 | 234.9 | 5.8 | 1.2 | 0.20 |
| | full | **246.5** | | | 239.2 | | | | |
| Lys | M-1 | 233.5 | 231.7 | 217.2 | 213.0 | 198.5 | 16.3 | 1.8 | 0.11 |
| | M-2 | 233.8 | 233.1 | 226.8 | 212.3 | 206.0 | 7.0 | 0.7 | 0.10 |
| | full | **230.5** | | | 212.5 | | | | |
| Arg | M-1 | 258.0 | 259.3 | 239.8 | 251.4 | 231.8 | 18.2 | −1.4 | −0.08 |
| | M-2 | 256.8 | 255.1 | 249.2 | 248.9 | 243.0 | 7.6 | 1.7 | 0.22 |
| | M-3 | 254.1 | 253.7 | 253.9 | 246.9 | 247.1 | 0.2 | 0.5 | 2.38 |
| | M-4 | 257.0 | 256.9 | 255.5 | 250.2 | 248.7 | 1.5 | 0.0 | 0.03 |
| | full | **253.9** | | | 250.5 | | | | |
| | | | | Tripeptides | | | | | |
| GEG | M-1 | 338.9 | 336.3 | 363.3 | 339.5 | 366.5 | −24.4 | 2.6 | −0.11 |
| | M-2 | 338.7 | 339.4 | 366.8 | 339.2 | 366.7 | −28.2 | −0.7 | 0.03 |
| | M-3 | 338.8 | 339.6 | 367.0 | 339.5 | 366.9 | −28.2 | −0.8 | 0.03 |
| | full | **338.2** | | | 340.2 | | | | |
| GHG | M-2 | 246.5 | 247.8 | 235.9 | 243.3 | 231.4 | 10.6 | −1.3 | −0.12 |
| | full | **248.5** | | | 242.1 | | | | |

[a] Separate values for B3LYP/6-31G(d) and DFTB applied to *real* and *model* systems are also listed. For each model, all energy calculations are performed at the ONIOM optimized geometry. The *real,B3LYP* and *real,DFTB* results therefore also vary with the ONIOM partition. For descriptions of the different computational models, see Figure 1. $\Delta S^{\rm high}$, $\Delta D$, and OES (ONIOM error score) are defined in section II.A. One-letter amino acid codes are used for the tripepeptides.

***Table 2.*** Proton Affinities (kcal/mol) for Two Tripeptides Using ONIOM(B3LYP:DFTB) with Different Size of the Basis Set in the B3LYP Calculation[a]

| real system | QM (high) basis | model system | B3LYP (real) | ONIOM | B3LYP (model) | DFTB (real) | DFTB (model) | $\Delta S^{\rm high}$ | $\Delta D$ | OES |
|---|---|---|---|---|---|---|---|---|---|---|
| GEG | 6-31G(d) | M-1 | 338.2 | 337.7 | 363.6 | 340.2 | 366.1 | −25.3 | 0.6 | −0.02 |
| | 6-31+G(d) | M-1 | 326.8 | 319.2 | 345.1 | 340.2 | 366.1 | −18.4 | 7.5 | −0.41 |
| | 6-311++G (2df, 2pd) | M-1 | 331.8 | 324.2 | 350.1 | 340.2 | 366.1 | −18.4 | 7.6 | −0.41 |
| GHG | 6-31G(d) | M-2 | 248.5 | 246.3 | 234.9 | 242.1 | 230.8 | 13.6 | 2.2 | 0.16 |
| | 6-31+G(d) | M-2 | 242.6 | 241.1 | 229.8 | 242.1 | 230.8 | 12.9 | 1.6 | 0.12 |
| | 6-311++G (2df, 2pd) | M-2 | 245.8 | 242.9 | 231.5 | 242.1 | 230.8 | 14.3 | 3.0 | 0.21 |

[a] The separate values for B3LYP and DFTB applied to *real* and *model* systems are also listed. Calculations are performed as single point on B3LYP/6-31G(d) optimized geometries. $\Delta S^{\rm high}$, $\Delta D$, and OES (ONIOM error score) are defined in section II.A. One-letter amino acid codes are used for the tripepeptides, and ONIOM cuts are shown in Figure 2.

tripeptide, the error of an imidazole model (M-2) is 10.6 kcal/mol, and the error is reduced to 1.3 kcal/mol in ONIOM. These results indicate that DFTB fairly well describes the effects of a DFT calculation with a double-ζ basis.

Extending the basis set from 6-31G(d) to 6-31+G(d), which includes diffuse functions, has large effects on the $\Delta S$ value (environmental effect) for the negative Gly−Glu−Gly system (from −25.3 to −18.4 kcal/mol), see Table 2. The DFTB $\Delta S$ value is closer to the B3LYP value without diffuse functions, and consequently the ONIOM error increases by 7 kcal/mol when going from the smallest to the largest basis set. Further increasing the basis to triple-ζ quality (6-311++(2df,2pd)) has a very limited effect.

The effect of diffuse functions is particularly large in the present case because the negatively charged *model* system is not well described with the small 6-31G(d) basis set; e.g., an acetate *model* system has occupied orbitals with positive eigenvalues. Adding diffuse functions thus stabilizes the acetate group relative to its base, leading to a significant basis set effect on the proton affinity. For better balanced systems, the effect of the environment is less basis-set-dependent, as seen by the results for the Gly−His−Gly (GHG) tripeptide in Table 2.

Most QM calculations of enzyme active sites use truncated amino acid models, and considering the large effects of the backbone on the calculated proton affinities (−18.4 kcal/

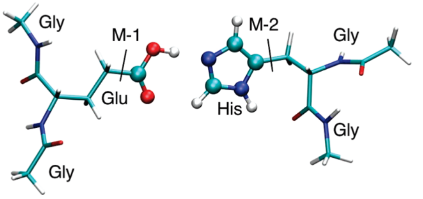


***Figure 2.*** Proton transfer between two tripeptides, Gly−Glu−Gly and Gly−His−Gly. Lines M-1 and M-2 illustrate the used *model* cuts (labels match the cuts made for the single amino acid systems in Figure 1). Atoms in ball-and-stick are included in the *model* system, while atoms in stick representation are part of the *real* system.

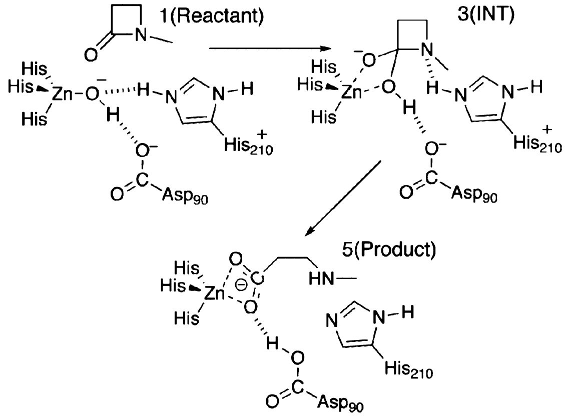
mol for Gly−Glu−Gly (GEG) with the largest basis, see Table 2) raises the question whether there is a significant error in the relative energies of proton transfer processes in active-site models. Comparing only the relative proton affinities of the *model*,*high* systems to those of the *real*,*high* systems for the tripeptides Gly−Glu−Gly and Gly−His−Gly (GHG) would give a difference in proton transfer energy of more than 30 kcal/mol. However, in the combined system, see Figure 2, the deviation between the *model*,*high* and the *real*,*high* calculation is only 1.8 kcal/mol, because that reaction does not include changes in the charge of the system. For this system, the use of ONIOM does not improve the result (deviation of −1.7 kcal/mol compared to the *model*,*high* value).

**B. Zn-$\beta$-lactamase−Various Active Site Models.** Enzyme catalysis is an attractive area for multiscale models, and the B3LYP/6-31G(d):DFTB combination is therefore evaluated for an enzymatic reaction, the hydrolysis of N-methylazetidinone in a mononuclear Zn-$\beta$-lactamase. Production of $\beta$-lactamase is considered the primary route in which bacteria acquire resistance to the common $\beta$-lactam antibiotics such as penicillins and cephalosporins. DFTB calculations of Zn active sites give relatively good agreement with B3LYP for distances and most reaction energies.[59] Modeling of substrate binding in a dinuclear Zn-$\beta$-lactamase shows relatively similar Zn−ligand distances for DFTB/MM simulations compared to B3LYP calculations (usually within 0.1 Å for reactants and intermediates).[60,61]

The reaction pathway and the initial coordinates were obtained from an active-site study by Diaz et al.[62] In their HF/6-31G(d) calculations, the reaction goes through five stationary points. As seen in Scheme 1, from 1(Reactant), a hydroxyl group bound to Zn performs a nucleophilic attack on the carbonyl of the four-membered $\beta$-lactam ring 2(TS). The ring is still intact in the tetrahedral intermediate 3(INT), and cleavage of the $\beta$-lactam C−N bond 4(TS) is initiated by proton transfer from His210 to the substrate nitrogen. The reaction also leads to a proton transfer from the Zn-coordinated hydroxo group to Asp90, leading to the final state 5(Product). At the B3LYP/6-31G(d) level, the reaction goes directly from 2(TS) to 5(Product) without any intermediates.[62]

To design a stringent test, a minimum-sized *model* system (model Z-1) was used, as shown in Figure 3. All ONIOM cuts have to be made through formal single bonds, and Asp90 is modeled as formic acid and the histidine residues as methylamine. This leaves 38 atoms in the *model* system (including link hydrogens). Calculations were initially performed at the HF optimized geometries because they cover a larger part of the reaction coordinate, see Figure 4 and Table S1 in Supporting Information.

***Scheme 1.*** Reaction Mechanism for a Mononuclear Zn-$\beta$-lactamase Adapted from Ref 62



The ONIOM energies for model Z-1 are improvements over the DFTB results (mean average deviation (MAD) for the 4 states in Figure 4 decreases to 4.3 kcal/mol from 10.6 kcal/mol), but only slightly better than the *model*,*B3LYP* calculations (MAD of 5.5 kcal/mol). The relatively large errors probably come from an inappropriate model partition, especially the use of methylamines as histidines in the *model* system. The 5-membered imidazole ring is conjugated while the *model* methylamine is not. Cuts in conjugated systems can be done in ONIOM, but they are in general more challenging. Similar partitions gave significant errors in the calculation of proton affinities, see Table 2. Errors can be systematically decreased by increasing the size of the *model* system. If the histidines that coordinate Zn are fully included in the *model* system (Z-2), the ONIOM error is reduced to 3.4 kcal/mol. Moving the proton-donating His210 into the *model* system (Z-3) further decreases the ONIOM deviation to 1.4 kcal/mol.

ONIOM deviations do not correlate with errors in the DFTB method itself. For 3(INT), the DFTB result is within 3 kcal/mol of the B3LYP value, but the ONIOM deviation in the small Z-1 model is 5.2 kcal/mol. For 5(Product), the DFTB deviation is 28 kcal/mol, but the ONIOM error is only 0.3 kcal/mol.

In the next step, we optimize the three stationary points on the B3LYP potential energy surface, 1(Reactant), 2(TS), and 5(Product), with both B3LYP and ONIOM(B3LYP: DFTB) using model Z-1; the results are shown in Figure 5. As shown in Table 3, two optimized geometries give only small energy difference up to 0.7 kcal/mol for either state at either level of calculations, indicating that the ONIOM geometry optimization is quite reliable for the reaction energetics.

**C. Reaction between Hydrogen Peroxide and a Nonheme Iron Catalase Mimic.** Redox-active transition-metal centers present special challenges in modeling due to the presence of nearly degenerate electronic states, which affects both the accuracy and the convergence properties of

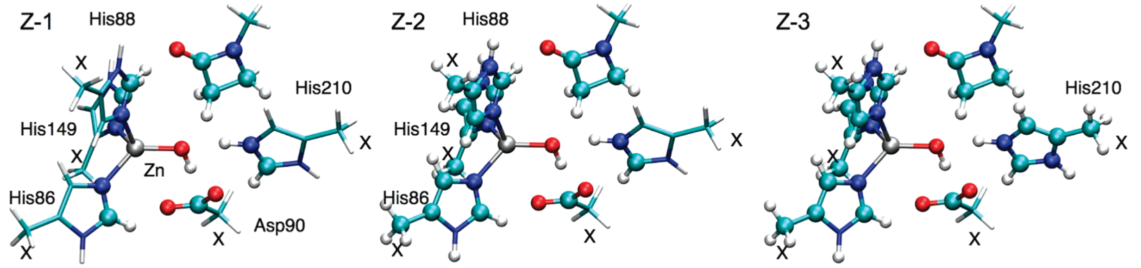


***Figure 3.*** ONIOM models (Z-1 to Z-3) of the active site in a mononuclear Zn-$\beta$-lactamase. Atoms in the *model* system are shown in ball-and-stick representation, while atoms in the *real* system are shown in stick representation. Atoms with coordinates frozen during optimizations are marked with X.

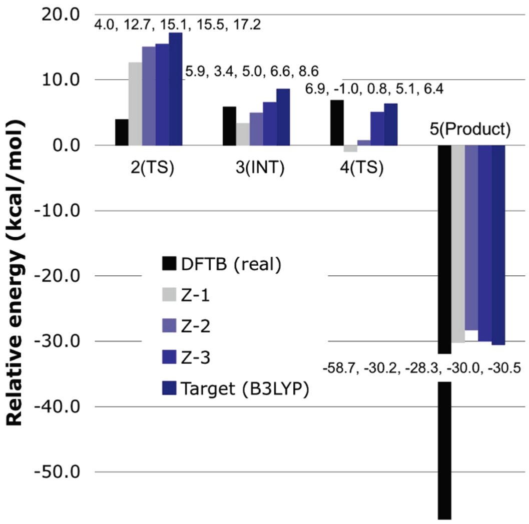


***Figure 4.*** DFTB, ONIOM(B3LYP/6-31G(d):DFTB), and B3LYP/6-31G(d) energies for hydrolysis of N-methylazetidinone in mononuclear Zn-$\beta$-lactamase in Scheme 1. Illustrations of ONIOM models Z-1 to Z-3 are given in Figure 3.

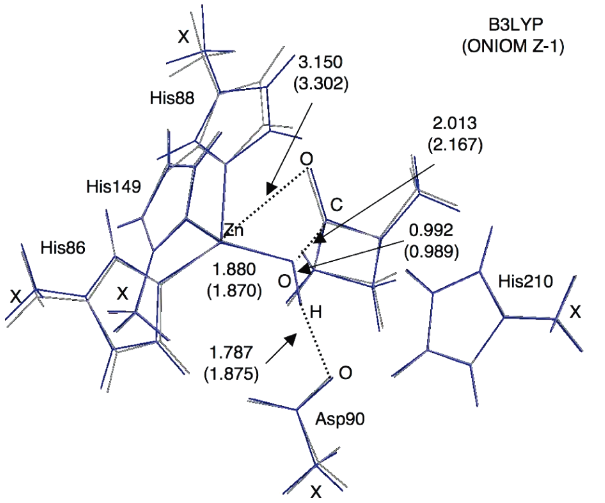


***Figure 5.*** Optimized structure of 2(TS) with B3LYP and ONIOM(B3LYP:DFTB) for model Z-1. Selected distances (in Å) are given at B3LYP/6-31G(d) level with ONIOM(B3LYP/6-31G(d):DFTB) results in parentheses. The total RMSD between the two structures is 0.124 Å.

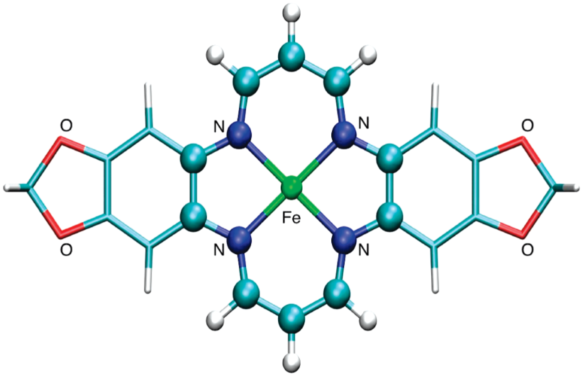


***Figure 6.*** ONIOM division for the dibenzotetraaza[14]annulene−$Fe^{III}$ complex. Atoms in the *model* system are shown in ball-and-stick representation, while atoms in the *real* system are shown in stick representation.

the electronic structure calculations. To extend the tests of the ONIOM(B3LYP:DFTB) approach to redox-active systems, we study the reaction between hydrogen peroxide and an inorganic catalase mimic, the dibenzotetraaza[14]annulene−$Fe^{III}$ complex ($[Fe(C_{24}H_{22}N_4O_4)]^+$),[63] see Figure 6. The ONIOM model is formed by truncating the conjugated ligand at bonds that can formally be assigned as single bonds.

QM:QM′ calculations of transition metal systems are challenging as three separate calculations of the transition metal system have to be performed. The three calculations must all converge to the same electronic state. Otherwise, the low-level method describes a different state from the high-level method, and the environmental effect in the *real*,*low* calculation becomes qualitatively incorrect. Despite these challenges, a previous test of a redox reaction in a nonheme transition metal enzyme gave reasonable results for the B3LYP:DFTB method.[48] With this in mind, it is of great interest to understand under what circumstances that QM:QM′ methods can be applied to transition metal systems.

The reaction mechanism of the catalase mimic has been previously studied by DFT methods.[64,65] The present investigation follows the reaction pathway in ref 64. A flaw in that study is that the additional axial (proximal) ligand was not taken into account.[65] For the present purpose, this is not critical because the comparison between methods should be valid even if the test system is not an ideal representation of an experimental situation.

According to ref 64, the reaction between hydrogen peroxide and the iron complex goes through nine stationary points and eventually leads to formation of water and an Fe(IV)−oxo species. The potential energy profiles from the full B3LYP calculation as well as ONIOM and *model*,*B3LYP* calculations are shown in Figure 7. Full results are given in Table 4. All calculations are performed on the quartet surface,

***Table 3.*** B3LYP/6-31G(d) and ONIOM (B3LYP/6-31G(d):DFTB) Energies (in kcal/mol) for the Hydrolysis of N-Methylazetidinone in Mononuclear Zn-$\beta$-lactamase at B3LYP and ONIOM Optimized Geometries

| model | state | B3LYP (real) | ONIOM | B3LYP (model) | DFTB (real) | DFT (model) | $\Delta S^{\mathrm{high}}$ | $\Delta D$ | OES |
|---|---|---|---|---|---|---|---|---|---|
| | | Geometries Obtained at the B3LYP/6-31G(d) Level | | | | | | | |
| Z-1 | 1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | NA | NA | NA |
| | 2(TS) | 16.2 | 10.1 | 12.3 | 2.2 | 4.4 | 3.9 | 6.1 | 1.57 |
| | 5(Prod) | −30.0 | −29.7 | −36.5 | −55.0 | −61.8 | 6.5 | −0.3 | −0.05 |
| | | Geometries Obtained at the ONIOM(B3LYP/6-31G(d):DFTB) Level | | | | | | | |
| Z-1 | 1 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | NA | NA | NA |
| | 2(TS) | 15.6 | 11.4 | 10.0 | 2.4 | 1.0 | 5.6 | 4.2 | 0.76 |
| | 5(Prod) | −30.2 | −31.1 | −35.7 | −55.0 | −59.6 | 5.6 | 0.9 | 0.16 |

and results are evaluated at the B3LYP/6-31G(d) geometries. Transition states have not been fully optimized; instead, the Fe−O, O−O, and O−H distances were taken from the study of Wang et al.[64] and kept frozen during the optimization. There are differences in the reported B3LYP potential energy surface in Figure 7 and the corresponding values in ref 64, and they come from differences in basis set between the two studies, as well as the neglect of zero-point energy in the present study.

The mean absolute deviation between B3LYP:DFTB and the target B3LYP results are 6.3 kcal/mol, and the maximum deviation is 13.2 kcal/mol, see Table 4. The performance of the B3LYP:DFTB method is much better than that of the stand-alone DFTB method, but the deviation is still too large to be acceptable in mechanistic studies. Adding the low-level correction in stationary points gives worse results than obtained by the *model*,*B3LYP* calculation, as seen from the large ONIOM error scores in Table 4.

To understand the ONIOM deviations, we compare the electronic structures obtained in the four separate calculations: *real*,*high*; *model*,*high*; *real*,*low*; and *model*,*low,* as shown in Table 5. At the first stationary point **1**, which is the catalyst before addition of hydrogen peroxide, the full B3LYP calculation (*real*,*high*) corresponds to an intermediate spin triplet Fe(II) which resulted from reduction of its formal Fe(III) state by an electron transfer from the annulene ligand. The reason for this electron transfer is probably the lack of the distal ligand.[65] The spin on iron couples ferromagnetically with the spin of the unpaired electron of the ligand to form a quartet state. The present assignment agrees with the results in ref 64 (unpaired electrons in $d_{z^2}$, $d_{xz}$, and the ligand $b_{1u}$ orbitals). The corresponding porphyrin−$Fe^{III}$ complex does not oxidize the ligand,[66] probably because the electron affinity is lower for the annulene ligand than for the porphyrin.[64] In B3LYP, the electronic structure of the *model* system is rather similar to that of the *real* system, see Table 5. The DFTB calculation of the *real* system gives a description more similar to an intermediate-spin quartet Fe(III) system; i.e., the oxidation of the ligand does not occur fully in DFTB. On the other hand, DFTB calculation for the *model* system converges to a high-spin quintet Fe(II) state that couples antiferromagnetically with the spin on the ligand, instead of an intermediate-spin Fe(II) state with ferromagnetic coupling to the ligand spin as in the B3LYP calculation, see Table 5. Thus, the ONIOM subtraction scheme provides the triplet Fe(II) state for the *model* system, but with the substituent effect evaluated using DFTB between the quartet Fe(III) *real* system and the quintet Fe(II) + ligand radical *model* system. This does not correspond exactly to the desired triplet Fe(II) state in the *real*,*B3LYP* calculations, and the QM:QM′ extrapolation scheme does not work as intended for this system.

The situation is similar for the stationary point **2**. Table 5 shows that *real*,*DFTB* and *model*,*B3LYP* give the quartet Fe(III) state, while *model*,*DFTB* gives the high-spin quintet Fe(II) state. Again the ONIOM scheme does not result in the correct description of the triplet Fe(II) state in the *real*,*B3LYP* calculations.

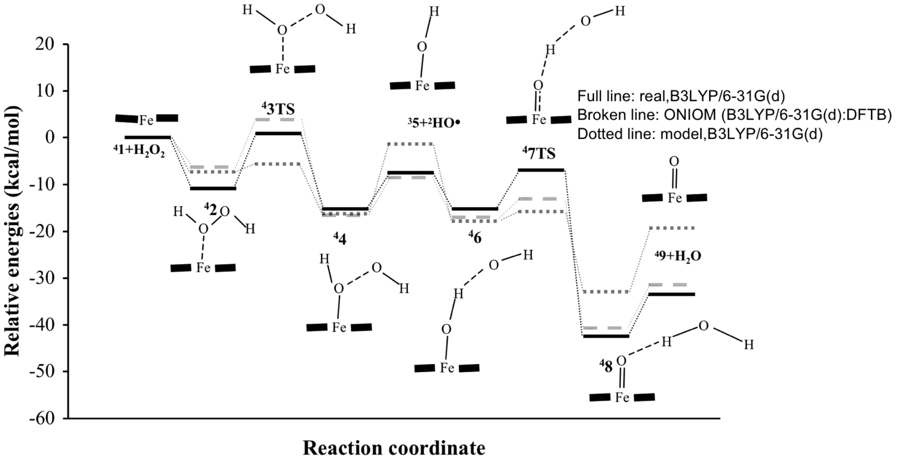


***Figure 7.*** B3LYP/6-31G(d) and ONIOM (B3LYP/6-31G(d):DFTB) potential energy profiles for the formation of a high-valent ferryl-oxo species in a catalase mimic. The ONIOM system is shown in Figure 6.

***Table 4.*** Relative Energies (in kcal/mol) for the Reaction between Hydrogen Peroxide and the Iron Complex Dibenzotetraaza[14]Annulene−Fe$^{III}$ Calculated Using ONIOM(B3LYP/6-31G(d):DFTB)[a]

| stationary point | B3LYP (real) | ONIOM | B3LYP (model) | DFTB (real) | DFTB (model) | $\Delta S^{high}$ | $\Delta D$ | OES |
|---|---|---|---|---|---|---|---|---|
| **1** | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | |
| **2** | −10.8 | −6.8 | −5.7 | −24.5 | −23.4 | −5.2 | −4.1 | 0.79 |
| **3** | 0.8 | −5.6 | 1.9 | −33.6 | −26.2 | −1.1 | 6.4 | −5.93 |
| **4** | −15.4 | −16.1 | −16.1 | −55.9 | −55.9 | 0.8 | 0.8 | 0.98 |
| **5** | −7.6 | −1.7 | −8.5 | −29.2 | −36.0 | 0.9 | −5.9 | −6.89 |
| **6** | −15.2 | −17.9 | −17.3 | −57.5 | −56.9 | 2.1 | 2.7 | 1.27 |
| **7** | −6.9 | −15.4 | −12.8 | −75.5 | −72.9 | 5.8 | 8.5 | 1.45 |
| **8** | −42.4 | −33.3 | −40.8 | −91.2 | −98.7 | −1.6 | −9.1 | 5.76 |
| **9** | −33.3 | −20.1 | −31.5 | −86.3 | −97.7 | −1.8 | −13.2 | 7.52 |
| mean absolute error | | | | | | 2.4 | 6.3 | |

[a] The separate values for B3LYP/6-31G(d) and DFTB applied to *real* and *model* systems are also listed. The different stationary points are shown in Figure 7. All calculations are performed at B3LYP/6-31G(d) optimized geometries.

***Table 5.*** Mulliken Spin Populations and Assigned Charges for the Formal Annulene−Fe$^{III}$ Complex[a]

| | | Mulliken spin population/assigned state label | | | |
|---|---|---|---|---|---|
| point | group | *real,B3LYP* | *real,DFTB* | *model,B3LYP* | *model,DFTB* |
| **1** | Fe | 2.16 /triplet Fe(II) | 3.39 /quartet Fe(III) | 2.00 /triplet Fe(II) | 4.10 /quintet Fe(II) |
| | ligand | 0.85/L$^{-1}$ | −0.39/L$^{-2}$ | 1.00/L$^{-1}$ | −1.10/L$^{-1}$ |
| **2** | Fe | 2.18/triplet Fe(II) | 3.33/quartet Fe(III) | 2.54/quartet Fe(III) | 4.01 /quintet Fe(II) |
| | ligand | 0.80/L$^{-1}$ | −0.35/L$^{-2}$ | 0.42/L$^{-2}$ | −1.03/L$^{-1}$ |
| | $H_2O_2$ | 0.03/0 | 0.03/0 | 0.04/0 | 0.02/0 |

[a] Only the first two stationary points in the reaction with hydrogen peroxide are listed. The assigned states are based on the number of unpaired electrons and should be taken as labels rather than exact assignment of states.

**D. The Acylation Process in Trypsin—Two- and Three-Layer Calculations.** The target of the ONIOM(QM:QM′) approach is to describe full enzymatic systems, but for systems with tens of thousands of atoms, the time required for the QM′ calculation would be very high. As an intermediate stage, the three-layer ONIOM(QM:QM′:MM) approach offers an efficient alternative. The three-layer QM:QM′:MM energy is obtained from five subcalculations:

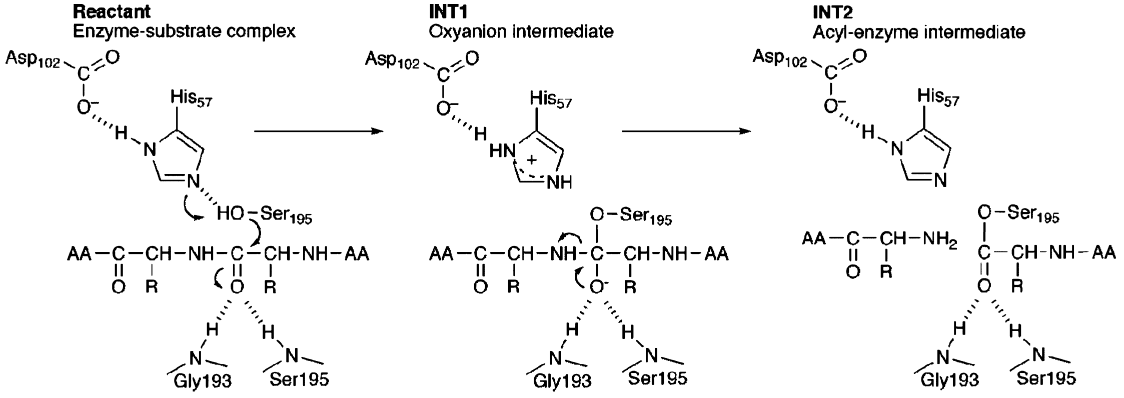
$$E^{\mathrm{ONIOM}} = E^{\mathrm{model,QM}} + E^{\mathrm{Intermediate,QM'}} - E^{\mathrm{model,QM'}} + E^{\mathrm{Real,MM}} - E^{\mathrm{Intermediate,MM}} \quad (7)$$

We illustrate the use of the three-layer approach for peptide cleavage in serine proteases. This reaction is one of the most well-known enzymatic reactions and appears in many biochemistry textbooks. Scheme 2 shows the proposed mechanism for the first half of the reaction, the acylation step.

An important motif in these enzymes is a conserved Ser−His−Asp triad, known as the catalytic triad. Ser195 performs a nucleophilic attack on the substrate peptide and transfers a proton to His57 (**INT1**). The negative Asp102 significantly stabilizes the proton transfer reaction by polarizing His57. The nucleophilic attack of serine leads to the formation of an oxyanion in the substrate peptide chain, and this species is stabilized by a second important catalytic motif, the “oxyanion hole”. This motif provides hydrogen bonds to the peptide carbonyl, interactions that increase in strength as the charge of the oxygen increases. In the next step, the peptide C−N bond breaks, and the newly formed N-terminal group accepts the proton from His57 (**INT2**). This completes the acylation part of the reaction.

*Active-Site QM:QM′ Models.* As the full enzyme cannot be benchmarked by QM calculations, we start with active-site models. To identify the effects of different catalytic motifs, we use three two-layer ONIOM systems (T-1 to T-3),

***Scheme 2.*** Textbook Mechanism for the Acylation Step in Serine Proteases



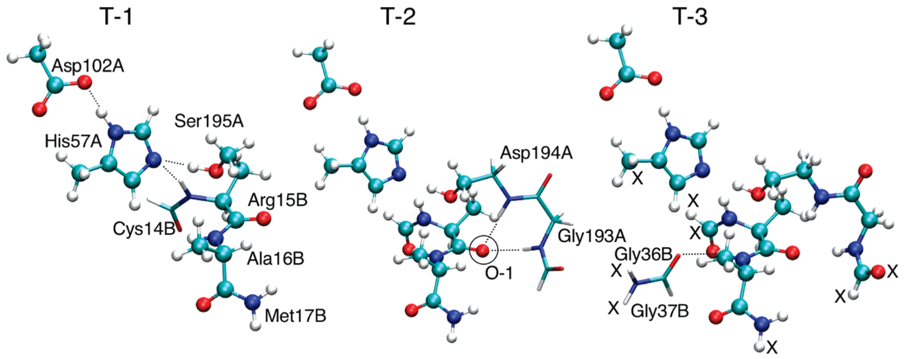


*Figure 8.* ONIOM divisions for an active-site model of trypsin. A and B in the amino acid labels refer to different peptide chains. Atoms in the *model* system are shown in ball-and-stick representation, while atoms in the *real* system are shown in stick representation. Hydrogen bonds are marked with dashed lines as they appear in the extended systems. Atoms with coordinates frozen at the X-ray structure are marked with **X** in model T-3.

*Table 6.* Reaction Energies (in kcal/mol, Relative to the Reactant) for the Acylation Process in Trypsin Calculated Using ONIOM B3LYP/6-31G(d):DFTB[a]

| model | state | B3LYP (real) | ONIOM | B3LYP (model) | DFTB (real) | DFTB (model) | $\Delta S^{\text{high}}$ | $\Delta D$ | OES |
|---|---|---|---|---|---|---|---|---|---|
| T-1 | **INT1** | 30.5 | 27.6 | 28.0 | 29.9 | 30.4 | 2.4 | 2.9 | 1.17 |
| | **INT2** | 2.8 | 4.6 | 8.1 | 2.0 | 5.6 | −5.3 | −1.7 | 0.32 |
| T-2 | **INT1** | 23.3 | 18.6 | 30.5 | 18.1 | 29.9 | −7.2 | 4.7 | −0.66 |
| | **INT2** | 11.6 | 8.4 | 2.8 | 7.6 | 2.0 | 8.8 | 3.3 | 0.37 |
| T-3 | **INT1** | 27.3 | 24.8 | 23.3 | 19.5 | 18.1 | 4.0 | 2.5 | 0.63 |
| | **INT2** | 12.6 | 11.7 | 11.6 | 7.6 | 7.6 | 0.9 | 0.9 | 0.92 |

[a] The separate values for B3LYP/6-31G(d) and DFTB applied to *real* and *model* systems are also listed. The different models are shown in Figure 8.

with different *model* and *real* selections, as shown in Figure 8. T-1 (47 atoms in the *model* system, 51 atoms in the *real* system) includes the catalytic triad and four amino acids of the substrate peptide, but one of the peptide units, Cys14B, is only part of the *real* system to test how the QM′ layer handles through-bond interactions. T-2 (51 atoms in the *model* system, 65 atoms in the *real* system) is created by adding the groups that stabilize the oxyanion hole. The environmental effect comes from two hydrogen bonds from the protein backbone. In DFTB, hydrogen bond distances are underestimated by about 0.1 Å on average, and hydrogen bonding energies are systematically underestimated by 1−2 kcal/mol.[51] To isolate the effect of this new motif, all atoms in the T-1 system are placed in the *model* system. ONIOM T-3 (65 atoms in the *model* system, 71 atoms in the *real* system) is created by adding a part of the backbone that forms a hydrogen bond with the amide NH group of the substrate peptide. Again, the new group is treated by QM′, while all the atoms from system T-2 are placed in the *model* system, see Figure 8.

Neglect of the surrounding protein makes the secondary structures unstable, and to avoid comparing different local minima, all calculations have been performed at the B3LYP/6-31G(d) optimized geometries of the full T-3 system. In this system, the relative energies of **INT1** and **INT2** are 27.3 and 12.6 kcal/mol, respectively. **INT1** is not a stationary point for the B3LYP/6-31G(d) optimization, and the structure is obtained by freezing the newly formed O−C distance at 1.513 Å (from ref 67). The relative energies are higher than expected for an enzymatic pathway and do not change significantly when applying a larger basis or a PCM solvent correction. It is possible that the reaction pathway is different in the present model than in the QM/MM free-energy perturbation calculations by Ishida and Kato,[67] but the exact reaction pathway is not critical for the comparison of QM and ONIOM(QM:QM′) results.

Results for the different B3LYP:DFTB models are given in Table 6. Results for AM1 and MM (Amber96) are given in the Supporting Information (Table S2). For **INT1**, the largest error (4.7 kcal/mol) comes from the treatment of the oxyanion hole with DFTB (T-2). At the B3LYP/6-31G(d) level, the addition of the oxyanion hole stabilizes **INT1** by 7.2 kcal/mol ($\Delta S^{\text{high}} = -7.2$ kcal/mol), which reflects one of the important enzymatic effects in the serine proteases. However, the DFTB layer gives a larger stabilization ($\Delta S^{\text{low}} = -11.9$ kcal/mol), leading to an ONIOM error of 4.7 kcal/mol.

The errors for the serine protease are much larger than those observed for a simple proton transfer reaction in which DFTB can well account for environmental effects of 23 kcal/mol with an error of only 1.2 kcal/mol.[47] The results illustrate that, as the environmental effect is calculated at the low-level only, not only must the low-level method be able to describe the electronic polarization effect but it must also be able to properly describe the changes in electronic structure of the reacting region. In DFTB, the electrostatic interactions are calculated using Mulliken charges,[43] and we therefore use Mulliken charges to discuss changes in charge distribution. In the reactant, DFTB assigns a Mulliken charge of −0.593 to the oxygen (O-1 in Figure 8), close to the value from B3LYP (−0.591), see Table 7. In **INT1**, DFTB assigns a much higher negative charge to the oxygen (−0.910) than the B3LYP calculation (−0.732), and it is not surprising that

***Table 7.*** Mulliken Charges for the Backbone Oxo Group of Arg15B That Becomes an Oxyanion in **INT1**[a]

| ONIOM | state | B3LYP (real) | DFTB (real) | B3LYP (model) | DFTB (model) |
|---|---|---|---|---|---|
| T-2 | Reac | −0.591 | −0.593 | −0.575 | −0.558 |
| | **INT1** | −0.732 | −0.910 | −0.730 | −0.915 |
| | **INT2** | −0.529 | −0.569 | −0.510 | −0.510 |

[a] The oxygen is labeled O-1 in Figure 8.

the stabilizing effect of the oxyanion hole is overestimated in the DFTB calculation. In **INT2**, the charge distributions are similar for B3LYP and DFTB, and ONIOM also gives smaller errors.

*Full Protein QM:QM′:MM Models.* To illustrate the applicability of the three-layer ONIOM (DFT:DFTB:MM) models for realistic system sizes, we designed an ONIOM model of trypsin that includes the entire protein and parts of the solvent shell. The initial structure for trypsin was taken from the PDB file 1TAW.[68] On the basis of a check of the structure with WHAT IF,[69] the terminal part of the Gln30 side chain was rotated. p$K_a$ values from PropKa suggested all His residues were singly protonated.[70] His40 and His91 were assigned as His$_\varepsilon$, and His57 was assigned as His$_\delta$. The protein was solvated in a flexible water box of approximate dimension 60 × 57 × 74 Å$^3$ using periodic boundary conditions. Seven chloride ions and four sodium ions were added in random positions in the solvent to achieve neutrality. With the pure MM method using the Amber96 force field, the system was minimized using a conjugate gradient method for 5000 steps, followed by 0.5 ns equilibration at 298 K with a constrained backbone and 1 ns without constraints, using the program NAMD 2.6.[71] After equilibration, an initial QM:MM *real* system was selected by including all protein atoms and all solvent molecules with at least one atom within 5 Å of the protein, totaling 6327 atoms. The selection includes one charged sodium atom from the solvation shell located in the vicinity of residue Phe34B at a position 16.3 Å from the peptide bond that is cleaved. Atoms more than 15 Å from any atom in the initial 88-atom *model* system were kept frozen in the optimizations. The 88-atom *model* system includes selected parts of residues Ala56, His57, Asp102, Cys191, Gly192, Gly193, Asp194, and Ser195 in chain A and residues Arg15, Ala16, Met17, Gly36, and Gly37 in chain B. Three stationary points are included in the test: the reactant, the oxyanion structure (**INT1**), and the intermediate with a cleaved C−N bond (**INT2**).

The 6327 atom protein structure was initially optimized at the B3LYP:MM level with a *model* QM system size of 88 atoms, see Figure 9. The QM selection is similar to the 71-atom active-site model (T-3 *real* system) but includes hydrogen-bonding groups whose mobility caused problems with geometry optimization in the active-site model. From the optimized reactant structure, the serine residue is moved toward the carbonyl of the peptide, until the O−C distance is 1.513 Å. This distance is frozen when the structure is reoptimized to form **INT1**. **INT2** is formed from **INT1** by elongating the peptide C−N bond and then freely optimizing all reaction coordinates. In studies of reaction mechanisms, the different stationary points should connect along the same potential energy surface to avoid effects from artificial changes in protein geometry during the optimizations. This typically requires several iterations between reactant and product until there no longer are any conformational changes in the surrounding protein. However, in the present study, we are mainly interested in the performance of different methods applied to the same stationary state, and going from reactant to product in a single step should still give relevant results.

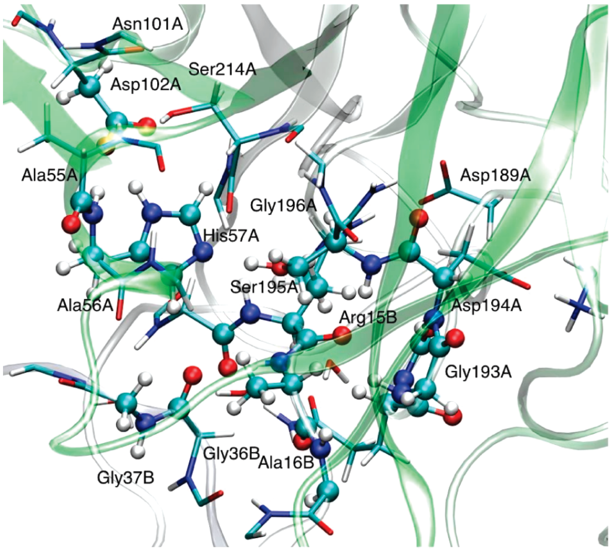


***Figure 9.*** ONIOM protein model of trypsin with 88 atoms in the *model* system (ball-and-stick representation), 215 atoms in the *intermediate* system (ball-and-stick and stick representations), and the *real* system (full protein) in cartoon representation.

The results of the new 88-atom *model* system are similar to those of the 71-atom active-site model. As an example, the relative energy of **INT1** is 27.0 and 27.3 kcal/mol, respectively, see Table 8. In all the calculations, the geometries optimized at the two-layer B3LYP/6-31G(d):MM level with 88 QM atoms are used.

First, as shown in Table 8, the QM system is extended to 215 atoms, and the performance of DFTB(215) and B3LYP(88):DFTB(215) (division D in Table 8) is compared to B3LYP(215) (division B in Table 8). The performance of the DFTB layer can be assessed by comparing the Δ*S* values for the 215-atom subsystem minus the 88-atom subsystem (column 7 in Table 8). The deviations are +2.4 kcal/mol for **INT1** and −4.8 kcal/mol for **INT2**.

Next, we compare the effects of using DFTB in the 795-atom subsystem instead of MM. In the present example, this DFTB layer is made up of the residues closest to the active site, but they can also be chosen on the basis of an evaluation of the protein effects in a previous QM:MM calculation. A comparison of the B3LYP(215):DFTB(793):MM (division A) and B3LYP(215):MM (division B) models shows that in **INT1** there is a relatively small effect, and a small difference (1.5 kcal/mol) between the two methods, see column 8 in Table 8. However, for **INT2**, the difference between the DFTB and the MM description of the environmental effect is big ∼12 kcal/mol. The situation for B3LYP(88):DFTB(793):MM (division C) and B3LYP(88):DFTB(215):MM (division D) models as well as for DFTB(793):MM (division E) and

***Table 8.*** The ONIOM Relative Energies and Their Components Δ*E* and Δ*S* (in kcal/mol, relative to the Reactant) for Different Two- and Three-Layer Models of Structures **INT1** and **INT2** of Trypsin[a]

| | systems and methods[b] | | | | Δ*E*[c] | Δ*S*[d] | | | |
|---|---|---|---|---|---|---|---|---|---|
| division | 88 | 215 | 793 | full | 88 | 215−88 | 793−215 | full−793 | ONIOM[e] |
| | | | | | **INT1** | | | | |
| A | B3LYP | B3LYP | DFTB | Amber | 27.0 | −4.3 | 1.2 | 5.0 | 29.0 |
| B | B3LYP | B3LYP | Amber | Amber | 27.0 | −4.3 | −0.3 | 5.0 | 27.5 |
| C | B3LYP | DFTB | DFTB | Amber | 27.0 | −1.8 | 1.2 | 5.0 | 31.4 |
| D | B3LYP | DFTB | Amber | Amber | 27.0 | −1.8 | −0.3 | 5.0 | 29.9 |
| E | DFTB | DFTB | DFTB | Amber | 31.3 | −1.8 | 1.2 | 5.0 | 35.7 |
| F | DFTB | DFTB | Amber | Amber | 31.3 | −1.8 | −0.3 | 5.0 | 34.2 |
| | | | | | **INT2** | | | | |
| A | B3LYP | B3LYP | DFTB | Amber | 14.2 | 0.6 | −0.8 | 2.7 | 16.7 |
| B | B3LYP | B3LYP | Amber | Amber | 14.2 | 0.6 | −12.4 | 2.7 | 5.1 |
| C | B3LYP | DFTB | DFTB | Amber | 14.2 | −4.3 | −0.8 | 2.7 | 11.8 |
| D | B3LYP | DFTB | Amber | Amber | 14.2 | −4.3 | −12.4 | 2.7 | 0.3 |
| E | DFTB | DFTB | DFTB | Amber | 12.5 | −4.3 | −0.8 | 2.7 | 10.1 |
| F | DFTB | DFTB | Amber | Amber | 12.5 | −4.3 | −12.4 | 2.7 | −1.4 |

[a] The full systems are formally divided into 88, 215, and 793 subsystems in Figure 9, and the method used for each subsystem is identified. Δ*E* is for the 88 atom system at the highest level, and Δ*S* is the difference between the two subsystems at the given level. [b] B3LYP is B3LYP/6-31G(d), and Amber is Amber96. [c] Δ*E* is the energy for the smallest 88 system at the highest level of the given division. [d] For instance, 1.2 in the row 4, column 8 represents $\Delta S^{793-215,\mathrm{DFTB}} = \Delta E^{793,\mathrm{DFTB}} - \Delta E^{215,\mathrm{DFTB}}$. [e] ONIOM energy is the sum of five contributions from column 6 to 9.

DFTB(215):MM (division F) is exactly the same as this effective is simply additive.

The large environmental effect at the MM level is partly due to changes in the structure, e.g., the orientation of a distant water (Wat39 in 1TAW numbering). These are simple artifacts of the optimization approach; i.e., different local MM minima are found for different intermediates. This would not be acceptable in a calculation of reaction energies, but as this is an ONIOM evaluation, no additional efforts were made to properly explore the MM energy landscape. From an ONIOM perspective, the most interesting data are the large differences between the DFTB and the MM description. Part of this difference probably comes from different relative energies for the artificial changes in protein geometry. Another effect is that the 793-atom DFTB subsystem allows transfer of charges and mutual polarization between the 215-atom subsystem and the 793-minus-215-atom layer, effects that are not included in the mechanical embedding MM method.[72]

## IV. Discussion

The simple calculations of proton affinities for amino acids illustrate important points about the B3LYP:DFTB models. The DFTB layer significantly improves the results compared to the *model* systems and gives good results also when the environmental effects are large. However, errors are not systematic; i.e., they do not decrease as the size of the *model* system increases. For the tripeptide systems, errors are 1−2 kcal/mol even for small *model* systems. As a comparison, errors of a carefully parametrized frozen orbital scheme are ∼1 kcal/mol for cuts in the backbone of the peptide chain.[57] For reactions where the environmental effect is basis-set-dependent, e.g., proton affinities of negative residues, the DFTB layer gives better results when combined with a double-ζ calculation, at which level DFTB is parametrized, compared to the triple-ζ calculation. These large effects are due to well-known problems of describing negative ions with an insufficient basis set., e.g., positive eigenvalues of occupied orbitals. However, for systems where the environmental effect is relatively independent of the basis set, e.g., proton affinities of positive residues, the DFTB layer works well independent of the basis set used in the QM calculation.

The QM:QM′ model of the nonheme catalase mimic illustrates an important point. The ONIOM method fails when QM and QM′ give different descriptions of the electronic state of the transition metal. This is unrelated to the *model* selection and can only be fixed by choosing more appropriate computational methods. Prior to the use of a QM:QM′ model, it is thus important to compare the electronic structure of the proposed *model* system with QM and QM′. Although B3LYP/6-31G(d) and DFTB give different energies for spin splitting energies of transition metal compounds, the two methods will often describe the same spin state as the multiplicity of the QM′ calculation is assigned at the start of the calculation. Problems with different electronic states should therefore only occur in systems with several electronic states for a given multiplicity. In the present example, the ligand radical could couple ferromagnetically with intermediate-spin iron or antiferromagnetically to high-spin iron. In our parametrization study, spin splittings of 13 Fe compounds had a mean average deviation of 15.2 kcal/mol and a maximum deviation of 34.3 kcal/mol when DFTB results were compared to B3LYP/SDD+6/31G(d).[48] This error is in the range of the difference between GGA and hybrid DFT functionals. However, in ONIOM, the requirement is only that the low-level method gives the same state as the high-level method, and out of the 67 complexes in ref 48, B3LYP and DFTB predict the same spin state in 52 cases.

Even in systems without open-shell species, the active-site B3LYP:DFTB models show some significant errors. In these tests, the reference *real*,*high* calculations must be affordable, and the tested systems are therefore always smaller, and the errors larger, than expected for normal applications. However, the trypsin calculations show that chemical accuracy (~1 kcal/mol) is not achieved even with relatively large QM models. For trypsin, these errors come from different descriptions of the charge distribution of the *model* system, e.g., of the oxyanion intermediate. This weakness of the mechanical embedding approach should be balanced against the savings in computational time.

When testing the applicability of a QM:QM′ model prior to its use, the first priority is to investigate if the reaction is qualitatively correctly described by the QM′ method. This can be done by comparing the electronic structure of the two methods, e.g., through population analysis, or by comparing changes in dipole moment. Another possibility is to make small ONIOM models and investigate the environmental effects of a single polarizing residue. If the effect of this residue is not qualitatively correct, the ONIOM model is not likely to give good results for any system selection.

More accurate results can be achieved by primarily using the QM:QM′ approach for optimizations. An important part of this paper is to show the possibility to fully optimize transition states with the QM:QM′ method. The energy can then be evaluated in a single-point calculation with a high-level QM method. For these calculations, the flexibility of the ONIOM method makes it possible to combine a larger basis set for the *model* system with a medium-sized basis set or low-cost method for the surrounding.

An interesting alternative to improve the description of electrostatic effects in QM:QM′ is the newly developed ONIOM(QM:QM′)−EE scheme.[73] In this scheme, the environmental effect at *real*,*low* is adjusted by the difference in response between QM and QM′ methods to a point charge environment obtained at the *real*,*QM′* level. QM:QM′−EE improves the DFT:HF results for several tested reactions.[74] As parts of the problem with the DFT:DFTB approach come from a different description of the reactive region, ONIOM−EE should improve the accuracy also with DFTB as the low-level method. However, the calculated environmental charges are geometry-dependent, and evaluation of QM:QM′−EE forces requires solutions of iterative coupled-perturbed equations, similar to coupled-perturbed Hartree−Fock.[73] This increases the cost of the *real*,*low* calculation approximately by a factor of 2, but it also means that not all methods can be directly used as low-layer methods. At the moment, QM:QM′−EE is only available with HF and DFT in the low layer.

We have also illustrated the use of a three-layer QM:QM′:MM model of an enzymatic system. Compared to standard QM/MM models, the scheme can be used in several different ways. In the trypsin model, parts of the MM system were replaced by DFTB, which in principle should lead to a better description of the environmental effect. It is also possible to reduce the size of the *model* system to allow for an improved description of the reactive region by correlated ab initio methods, e.g., CCSD(T) or MRCI.

Optimization of large three-layer models is still difficult. QM:MM optimizations are performed with microiterations where the MM system is optimized using a first-order algorithm. If a DFTB layer of ~1000 atoms is optimized together with the MM part, several hundred QM′ energy and gradient evaluations will be performed for each QM iteration. The computational time for the QM′ calculations will then become higher than the time required for the QM calculation. A second alternative is to optimize the QM′ layer together with the QM region in a second-order algorithm. The drawbacks are that the optimization algorithm becomes more costly, and more importantly, the number of macroiterations increases. This leads to an increase in the number of QM calculations required to reach geometry convergence. Higher efficiency might be reached with a hybrid technique that uses three different optimization levels, but such a scheme has yet to be implemented.

## V. Conclusions

The ONIOM(QM:QM′) scheme with mechanical embedding is cost-efficient as it only requires a single QM evaluation at each geometry. The drawback is that all environmental effects are evaluated at the low (QM′) level, and the accuracy of the scheme depends on how well the low-level QM′ method describes the environmental effects and the changes in electron density during the reaction.

To illustrate the advantages and limitations of this method, we have applied ONIOM(B3LYP:DFTB) and three-layer ONIOM(B3LYP:DFT:MM) combinations to models of enzymes and enzyme mimics. Although the DFTB layer reduces a large part of the error in the underlying *model* calculations, remaining errors of several kilocalories per mole are not uncommon. The polarization effects are fairly well described using DFTB, but the QM and QM′ methods do not always describe the same electronic state throughout the reaction, causing some difficulties.

Use of the QM:QM′ model requires an in-depth understanding of both the QM and the QM′ method, and the applicability of the QM:QM′ scheme must be carefully investigated in each application. Separate calculations of the electronic structure using both QM and QM′ methods as well as a test of small ONIOM systems offer a reasonable way to test a QM:QM′ scheme without performing extensive benchmark tests.

The good performance of the DFTB method for geometries, together with the possibility to optimize transition states with the QM:QM′ scheme, makes it an excellent tool for exploration of geometries. Accurate energies can then be obtained by single-point calculations using a high-level method.

**Acknowledgment.** One of the authors (M.L.) acknowledges a Fukui Institute for Fundamental Chemistry Fellowship. The present work was in part supported by a CREST (Core Research for Evolutional Science and Technology) grant in the Area of High Performance Computing for Multiscale and Multiphysics Phenomena from the Japan Science and Technology Agency (JST). The computational resources at the Research Center for Computational Science at Institute for Molecular Science are gratefully acknowledged.

**Supporting Information Available:** Table S1 listing ONIOM energies for hydrolysis of N-methylazetidinone in mononuclear Zn-$\beta$-lactamase. Table S2 listing ONIOM energies for active-site model of peptide hydrolysis in trypsin.

PDB files for trypsin stationary points. Gaussian input files with ONIOM layer selections (88, 215, and 793 atoms) for trypsin. This information is available free of charge via the Internet at http://pubs.acs.org/.

## References

(1) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227–249.

(2) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1986**, *7*, 718–730.

(3) Field, M. J.; Bash, P. A.; Karplus, M. *J. Comput. Chem.* **1990**, *11*, 700–733.

(4) Lin, H.; Truhlar, D. G. *Theor. Chem. Acc.* **2007**, *117*, 185–199.

(5) Senn, H. M.; Thiel, W. In *Atomistic Approaches in Modern Biology: From Quantum Chemistry to Molecular Simulations*; Springer: New York, 2007; Vol. 268, pp 173−290.

(6) Maseras, F.; Morokuma, K. *J. Comput. Chem.* **1995**, *16*, 1170–1179.

(7) Humbel, S.; Sieber, S.; Morokuma, K. *J. Chem. Phys.* **1996**, *105*, 1959–1967.

(8) Matsubara, T.; Maseras, F.; Koga, N.; Morokuma, K. *J. Phys. Chem.* **1996**, *100*, 2573–2580.

(9) Svensson, M.; Humbel, S.; Froese, R. D. J.; Matsubara, T.; Sieber, S.; Morokuma, K. *J. Phys. Chem.* **1996**, *100*, 19357–19363.

(10) Dapprich, S.; Komaromi, I.; Byun, K. S.; Morokuma, K.; Frisch, M. J. *J. Mol. Struct. Theochem* **1999**, *462*, 1–21.

(11) Vreven, T.; Morokuma, K. *J. Comput. Chem.* **2000**, *21*, 1419–1432.

(12) Vreven, T.; Byun, K. S.; Komaromi, I.; Dapprich, S.; Montgomery, J. A.; Morokuma, K.; Frisch, M. J. *J. Chem. Theory Comput.* **2006**, *2*, 815–826.

(13) Derat, E.; Bouquant, J.; Humbel, S. *J. Mol. Struct. Theochem* **2003**, *632*, 61–69.

(14) Morokuma, K.; Wang, Q. F.; Vreven, T. *J. Chem. Theory Comput.* **2006**, *2*, 1317–1324.

(15) Froese, R. D. J.; Morokuma, K. *Chem. Phys. Lett.* **1999**, *305*, 419–424.

(16) Vreven, T.; Morokuma, K. *J. Phys. Chem. A* **2002**, *106*, 6167–6170.

(17) Morokuma, K. *Bull. Chem. Soc. Jpn.* **2007**, *80*, 2247–2261.

(18) Lundberg, M.; Morokuma, K. *In Multi-scale Quantum Models for Biocatalysis: Modern Techniques and Applications*, Eds. Lee, T.-S. York, D. M., Springer Verlag 2009.

(19) Stern, H. A.; Kaminski, G. A.; Banks, J. L.; Zhou, R. H.; Berne, B. J.; Friesner, R. A. *J. Phys. Chem. B* **1999**, *103*, 4730–4737.

(20) Stern, H. A.; Rittner, F.; Berne, B. J.; Friesner, R. A. *J. Chem. Phys.* **2001**, *115*, 2237–2251.

(21) Morales, J.; Martinez, T. J. *J. Phys. Chem. A* **2004**, *108*, 3076–3084.

(22) Gascon, J. A.; Leung, S. S. F.; Batista, E. R.; Batista, V. S. *J. Chem. Theory Comput.* **2006**, *2*, 175–186.

(23) Chen, J. H.; Martinez, T. J. *Chem. Phys. Lett.* **2007**, *438*, 315–320.

(24) Chen, J. H.; Hundertmark, D.; Martinez, T. J. *J. Chem. Phys.* **2008**, *129*, 214113.

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

(26) Ren, P. Y.; Ponder, J. W. *J. Phys. Chem. B* **2003**, *107*, 5933–5947.

(27) Xie, W. S.; Song, L. C.; Truhlar, D. G.; Gao, J. L. *J. Phys. Chem. B* **2008**, *112*, 14124–14131.

(28) Xie, W. S.; Gao, J. L. *J. Chem. Theory Comput.* **2007**, *3*, 1890–1900.

(29) Zhang, Y.; Lin, H. *J. Chem. Theory Comput.* **2008**, *4*, 414–425.

(30) Yang, W. T.; Lee, T. S. *J. Chem. Phys.* **1995**, *103*, 5674–5678.

(31) Wesolowski, T. A.; Weber, J. *Chem. Phys. Lett.* **1996**, *248*, 71–76.

(32) Nakano, T.; Kaminuma, T.; Sato, T.; Akiyama, Y.; Uebayasi, M.; Kitaura, K. *Chem. Phys. Lett.* **2000**, *318*, 614–618.

(33) Cortona, P. *Phys. Rev. B* **1991**, *44*, 8454–8458.

(34) Henderson, T. M. *J. Chem. Phys.* **2006**, *125*, 014105.

(35) Huang, P.; Carter, E. A. *J. Chem. Phys.* **2006**, *125*, 084102.

(36) Huang, P.; Carter, E. A. *Annu. Rev. Phys. Chem.* **2008**, *59*, 261–290.

(37) Gogonea, V.; Westerhoff, L. M.; Merz, K. M. *J. Chem. Phys.* **2000**, *113*, 5604–5613.

(38) Cui, Q.; Guo, H.; Karplus, M. *J. Chem. Phys.* **2002**, *117*, 5617–5631.

(39) Svensson, M.; Humbel, S.; Morokuma, K. *J. Chem. Phys.* **1996**, *105*, 3654–3661.

(40) Vreven, T.; Morokuma, K. *J. Chem. Phys.* **1999**, *111*, 8799–8803.

(41) Tschumper, G. S.; Morokuma, K. *J. Mol. Struct. Theochem* **2002**, *592*, 137–147.

(42) Porezag, D.; Frauenheim, T.; Kohler, T.; Seifert, G.; Kaschner, R. *Phys. Rev. B* **1995**, *51*, 12947–12957.

(43) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260–7268.

(44) Kohler, C.; Seifert, G.; Gerstmann, U.; Elstner, M.; Overhof, H.; Frauenheim, T. *Phys. Chem. Chem. Phys.* **2001**, *3*, 5109–5114.

(45) Frauenheim, T.; Seifert, G.; Elstner, M.; Hajnal, Z.; Jungnickel, G.; Porezag, D.; Suhai, S.; Scholz, R. *Phys. Status Solidi B* **2000**, *217*, 41–62.

(46) Zheng, G. S.; Irle, S.; Morokuma, K. *Chem. Phys. Lett.* **2005**, *412*, 210–216.

(47) Zheng, G. S.; Lundberg, M.; Jakowski, J.; Vreven, T.; Frisch, M. J.; Morokuma, K. *Int. J. Quantum Chem.* **2009**, *184*, 1–1854.

(48) Zheng, G. S.; Witek, H. A.; Bobadova-Parvanova, P.; Irle, S.; Musaev, D. G.; Prabhakar, R.; Morokuma, K.; Lundberg, M.; Elstner, M.; Köhler, C.; Frauenheim, T. *J. Chem. Theory Comput.* **2007**, *3*, 1349–1367.

(49) Iordanov, T. D. *THEOCHEM* **2008**, *850*, 152–159.

(50) Elstner, M.; Jalkanen, K. J.; Knapp-Mohammady, M.; Frauenheim, T.; Suhai, S. *Chem. Phys.* **2001**, *263*, 203–219.

(51) Elstner, M. *Theor. Chem. Acc.* **2006**, *116*, 316–325.

(52) Otte, N.; Scholten, M.; Thiel, W. *J. Phys. Chem. A* **2007**, *111*, 5751–5755.

(53) Nakatani, N.; Hasegawa, J. Y.; Nakatsuji, H. *J. Am. Chem. Soc.* **2007**, *129*, 8756–8765.

(54) Wanko, M.; Hoffmann, M.; Frauenheim, T.; Elstner, M. *J. Phys. Chem. B* **2008**, *112*, 11462–11467.

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

(56) Siegbahn, P. E. M. *J. Comput. Chem.* **2001**, *22*, 1634–1645.

(57) Murphy, R. B.; Philipp, D. M.; Friesner, R. A. *J. Comput. Chem.* **2000**, *21*, 1442–1457.

(58) Zhou, H. Y.; Tajkhorshid, E.; Frauenheim, T.; Suhai, S.; Elstner, M. *Chem. Phys.* **2002**, *277*, 91–103.

(59) Elstner, M.; Cui, Q.; Munih, P.; Kaxiras, E.; Frauenheim, T.; Karplus, M. *J. Comput. Chem.* **2003**, *24*, 565–581.

(60) Xu, D. G.; Guo, H.; Cui, Q. *J. Phys. Chem. A* **2007**, *111*, 5630–5636.

(61) Xu, D.; Guo, H.; Cui, G. *J. Am. Chem. Soc.* **2007**, *129*, 10814–10822.

(62) Diaz, N.; Suarez, D.; Merz, K. M. *J. Am. Chem. Soc.* **2001**, *123*, 9867–9879.

(63) Paschke, J.; Kirsch, N.; Korth, H. G.; de Groot, H.; Sustmann, R. *J. Am. Chem. Soc.* **2001**, *123*, 11099–11100.

(64) Wang, X.; Li, S. H.; Jiang, Y. S. *Inorg. Chem.* **2004**, *43*, 6479–6489.

(65) Sicking, W.; Korth, H. G.; Jansen, G.; de Groot, H.; Sustmann, R. *Chem.—Eur. J.* **2007**, *13*, 4230–4245.

(66) Oliveira, K. M. T.; Trsic, M. *THEOCHEM* **2001**, *539*, 107–117.

(67) Ishida, T.; Kato, S. *J. Am. Chem. Soc.* **2003**, *125*, 12035–12048.

(68) Scheidig, A. J.; Hynes, T. R.; Pelletier, L. A.; Wells, J. A.; Kossiakoff, A. A. *Protein Sci.* **1997**, *6*, 1806–1824.

(69) Vriend, G. *J. Mol. Graphics* **1990**, *8*, 52–56.

(70) Li, H.; Robertson, A. D.; Jensen, J. H. *Proteins: Struct., Funct., Bioinf.* **2005**, *61*, 704–721.

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

(72) Lundberg, M.; Kawatsu, T.; Vreven, T.; Frisch, M. J.; Morokuma, K. *J. Chem. Theory Comput.* **2009**, *5*, 222–234.

(73) Hratchian, H. P.; Parandekar, P. V.; Raghavachari, K.; Frisch, M. J.; Vreven, T. *J. Chem. Phys.* **2008**, *128*, 034107.

(74) Parandekar, P. V.; Hratchian, H. P.; Raghavachari, K. *J. Chem. Phys.* **2008**, *129*, 145101.

CT100029P

JCTC Journal of Chemical Theory and Computation

# Fragment-Molecular-Orbital-Method-Based *ab Initio* NMR Chemical-Shift Calculations for Large Molecular Systems

Qi Gao,[†,‡] Satoshi Yokojima,*[,†,¶] Dmitri G. Fedorov,[§] Kazuo Kitaura,[§,||] Minoru Sakurai,[‡] and Shinichiro Nakamura*[,†,‡,¶]

*Mitsubishi Chemical Group Science and Technology Research Center, Inc., 1000 Kamochida-cho, Aoba-ku, Yokohama 227-8502, Japan, Center for Biological Resources and Informatics, Tokyo Institute of Technology, 4259 Nagatsuda-cho, Midori-ku, Yokohama 226-8501, Japan, The KAITEKI Institute, Inc. 14-1, Shiba 4-chome, Minato-ku, Tokyo 108-0014, Japan, RICS, National Institute of Advanced Industrial Science and Technology (AIST), 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan, and Graduate School of Pharmaceutical Sciences, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan*

Received October 6, 2009

**Abstract:** An *ab initio* computational method, based on the fragment molecular orbital (FMO) method, for calculating nuclear magnetic resonance (NMR) chemical shifts has been developed by introducing the concept of a merged fragment with a cutoff distance. Using point charges or density based on electrostatic potential obtained from FMO calculations, the NMR calculations (GIAO and CSGT) with the 6-31G(d) and 6-311G(d,p) basis sets were performed on $\alpha$-helix and $\beta$-sheet polypeptides. The cutoff distance defines the optimal merged fragment size for NMR calculations. This method accurately reproduces electrostatic effects and magnetic susceptibilities. The chemical shifts determined with a cutoff distance not shorter than 8 Å for both $\alpha$-helix and $\beta$-sheet polypeptides agree well with those calculated by conventional *ab initio* NMR calculations.

## 1. Introduction

In the fields of chemical and biological research, nuclear magnetic resonance (NMR) spectroscopy is one of the most valuable experimental methods for determining chemical and structural properties. Since NMR chemical shifts are very sensitive to changes in the molecular structure, they are widely used for monitoring the surrounding environment of individual atoms. However, in the case of large biological molecules, the process of obtaining the required signal-to-noise ratio (SNR) when measuring NMR spectra is time-consuming.

Several theoretical methods for computing NMR chemical shifts of *small* organic molecules have been established. These methods can be divided into two primary categories. The first are *ab initio* methods dealing with gauge-dependence problems such as the "gauge-including atomic orbital" (GIAO)[1,2] and the "continuous set of gauge transformations" (CSGT).[3] Those methods can reliably be applied to structures of organic and inorganic compounds using appropriate basis sets. However, when it comes to large molecules such as proteins, the computational cost becomes quite large, so not so many applications of *ab initio* methods have been reported.[4] The second category is empirical methods that predict NMR chemical shifts in a very short time using a large database. Nonetheless, because these methods have not been parametrized to reproduce NMR chemical shifts of an arbitrary molecule, they fail to predict chemical shifts of some new specific compounds.

* To whom corresponding should be addressed. Tel: +81-45-963-3833 (S.Y.), +81-45-963-3265 (S.N.). Fax: +81-45-963-3835 (S.Y.), +81-45-963-3835 (S.N.). E-mail: yokojima@rsi.co.jp (S.Y.), shindon@rsi.co.jp (S.N.).

[†] Mitsubishi Chemical Group Science and Technology Research Center, Inc.

[¶] The KAITEKI Institute, Inc.

[‡] Tokyo Institute of Technology.

[§] AIST.

[||] Kyoto University.

10.1021/ct100006n © 2010 American Chemical Society
Published on Web 03/01/2010

The hybrid quantum-mechanical/molecular-mechanical (QM/MM) method[5,6] is a computational tool that can be used to predict the NMR chemical shifts in large molecules. In the QM/MM method, a part for calculating NMR chemical shifts is described with QM methods, whereas the remaining part is described with MM methods. Cui and Karplus proposed a method for calculating chemical shifts using the GIAO in the QM/MM framework.[7] After carefully comparing the results of the QM/MM model to those of full QM calculations, they concluded that the former can provide good descriptions of environmental effects on chemical shifts. The error compared to full QM calculations is 1−2 ppm for heavy atoms when the distance between the atom of interest in the QM part and the MM part is more than 2.5 Å. The importance of the contributions of the Pauli repulsion and magnetic susceptibility is also evident from their studies. If these contributions are neglected at distances of less than 2.5 Å,[7] large errors occur. Similar results have also been reported by Ishida,[8] He et al.,[9] and Sebastiani et al.[10–12] For example, Sebastiani and Rothlisberger[10,11] have modified the standard QM/MM interaction potential by including the Pauli repulsion explicitly in the QM/MM interaction potential. They concluded that the modified QM/MM interaction potential can reproduce full QM results.

In contrast with these QM/MM NMR studies in which the interactions between QM and MM regions have been handled as described by Field et al.,[13] a study by Gascón et al.[14,15] used the ONIOM[16] electronic-embedding method to calculate the NMR data of large molecules, and their calculation results agreed well with experimental measurements. Hall et al.[17] have recently used a three-layer ONIOM (B3LYP:HF:AMBER) scheme to calculate the NMR chemical shifts of the retinal chromophore in rhodopsin.

In addition to the partial quantum-mechanical description, there are some fragment-based methods where the whole system is described quantum mechanically.[18–25] The fragment molecular orbital (FMO) method[26–31] is one of such fragment-based methods applied to *ab initio* calculations[32–36] of large molecules. Sekino et al.[37] have developed the FMO2-NMR method, where the FMO many-body expansion has been applied to the computation of chemical shifts.

We recently developed an *ab initio* method[38] based on FMO1 for computing the NMR response in protein systems. This method involves inexpensive single-fragment NMR calculations including the electrostatic effects from other fragments. Our method[38] of computing NMR chemical shifts consists of two parts. First, a molecular system is divided into various fragments, and the electron densities of these fragments are obtained by carrying out FMO calculations. Second, for the neighboring dimer fragments in the electrostatic potential (ESP) of other fragments, nuclear magnetic shielding constants are computed with GIAO and/or CSGT methods. Here, the chemical shift, $\delta$, is related to the nuclear magnetic shielding tensor, $\sigma$, by a reference standard, $\sigma^0$, as[39] $\delta = (\sigma^0 - \sigma^{\text{iso}})/(1 - \sigma^0) \times 10^6 \approx (\sigma^0 - \sigma^{\text{iso}}) \times 10^6$, where the absolute isotropic shielding constant, $\sigma^{\text{iso}} = \text{Tr}(\sigma)/3$.

Although this method can reasonably reproduce experimental and conventional *ab initio* calculated chemical shifts, errors of 2 ppm for $^{13}$C, 4.5 ppm for $^{15}$N, and 0.8 ppm for $^1$H are still larger than the desired level, namely, below 0.5 ppm for heavy atoms and 0.1 ppm for hydrogen atoms.[38] Furthermore, these calculations have been less accurate in predicting anisotropic shielding constants.[38] We previously used our method[38] to calculate chemical shifts of a $\beta$ sheet.[40] However, the errors were even larger than those stated above because of a lack of accuracy in reproducing the electrostatic effects and magnetic susceptibilities.

Previous quantum chemical and experimental studies[41–44] on NMR chemical shifts in various chemical environments have indicated that the electrostatic effect and magnetic susceptibility arise not only from covalent-bond interactions but also from weaker interactions such as hydrogen bonds, the susceptibility of conjugated carbon groups, and the polarization of the surrounding environment. These studies have emphasized that, to accurately reproduce the electron distributions around atoms, the inclusion of these interactions is crucial.

Other sources of error due to our method[38] may arise from border atoms, i.e., atoms on the border of two fragments. We have analyzed the influence of border atoms on the ESP around them. When the positions of border atoms are far enough away from the atom of interest, our method[38] reproduced the ESP of the conventional *ab initio* method. However, when these distances are shorter, the method yields relatively inaccurate ESP, resulting in large errors in chemical shifts.

The purpose of the present study was to develop a new method—using a new cutoff-distance-based fragmentation scheme—that can accurately reproduce conventional *ab initio* isotropic and anisotropic shielding constants. This method adopts the effective framework of the FMO1 method, which is conceptually similar to the multilayer structure of QM/MM. Unlike the QM/MM method, the effect of the MM part is described by the ESP derived from quantum mechanics. The main difference of the cutoff-distance-based method from our previous method for computing nuclear magnetic shielding constants from the electronic structure of neighboring dimer fragments[38] is that a cutoff distance to define the size of merged fragments is introduced to calculate nuclear magnetic shielding constants. The density matrix of a merged fragment describes the electrostatic response better than that of a dimer fragment. In addition, the cumbersome use of the two fragmentation schemes in our previous method is eliminated. Using the cutoff-distance-based method, we assessed the dependence of nuclear magnetic shielding constants on cutoff distance, basis sets, and polypeptide structures. Excellent agreement between the shielding constants of the cutoff-distance-based method with those of the conventional *ab initio* method is attained when the cutoff distance is ≥8 Å. The isotropic shielding constants of all the tested heavy atoms ($^{13}$C and $^{15}$N) and $^1$H atoms obtained by the cutoff-distance-based method with a cutoff distance of 8 Å reproduce the values obtained with the conventional *ab initio* method (CSGT) within 0.24 ppm and 0.11 ppm, respectively.

## 2. Theory

**2.1. FMO Method.** The FMO method[28,45] developed by Kitaura et al. is designed to calculate the electronic state of large systems within the *ab initio* framework. First, a large system is divided into $N$ fragments (monomers). The total energy is then given by

$$E = \sum_{I}^{N} E_I + \sum_{I>J}^{N} (E_{IJ} - E_I - E_J) \tag{1}$$

where energy $E_I$ of monomer $I$ and energy $E_{IJ}$ of dimer $IJ$ (a pair of monomers $I$ and $J$) are summed. At the RHF level, the contributions of the monomer and dimer to the total energy can be calculated by

$$\tilde{\boldsymbol{F}}^x \mathbf{C}^x = \boldsymbol{S}^x \boldsymbol{C}^x \tilde{\boldsymbol{\varepsilon}}^x \tag{2}$$

$$\tilde{\boldsymbol{F}}^x = \tilde{\boldsymbol{H}}^x + \boldsymbol{G}^x \tag{3}$$

$$\tilde{H}^x_{\mu\nu} = H^x_{\mu\nu} + V^x_{\mu\nu} + \gamma \sum_{i \in x} P^i_{\mu\nu} \tag{4}$$

$$P^i_{\mu\nu} = \langle \mu | \varphi^h_i \rangle \langle \varphi^h_i | \nu \rangle \tag{5}$$

$$V^x_{\mu\nu} = \sum_{K \neq x} \left\{ \sum_{A \in K} \left\langle \mu \left| -\frac{1}{4\pi\varepsilon_0} \frac{Z_A e^2}{|\boldsymbol{r} - \boldsymbol{R}_A|} \right| \nu \right\rangle + \sum_{\rho\sigma \in K} D^K_{\rho\sigma} (\mu\nu|\rho\sigma) \right\} \tag{6}$$

$$(\mu\nu|\rho\sigma) = \frac{e^2}{4\pi\varepsilon_0} \int \mathrm{d}\boldsymbol{r}_1 \, \mathrm{d}\boldsymbol{r}_2 \, \chi_\mu(\boldsymbol{r}_1) \, \chi_\nu(\boldsymbol{r}_1) |\boldsymbol{r}_1 - \boldsymbol{r}_2{}^{-1} \chi_\rho(\boldsymbol{r}_2) \, \chi_\sigma(\boldsymbol{r}_2) \tag{7}$$

$$E_x = \frac{1}{2} \mathrm{Tr}\{\boldsymbol{D}^x (\tilde{\boldsymbol{H}}^x + \tilde{\boldsymbol{F}}^x)\} + E^x_{\mathrm{Nuc}} \tag{8}$$

where superscript $x$ represents a monomer ($x = I$) or a dimer ($x = IJ$); $\mu$, $\nu$, $\rho$, and $\sigma$ denote atomic orbitals; $Z_A$ and $\boldsymbol{R}_A$ correspond to the charge and the position of atom $A$; and $\boldsymbol{D}^K$ is the density matrix of fragment $K$. The Fock matrix, $\tilde{\boldsymbol{F}}^x$, consists of the sum of one-electron term $\tilde{H}^x_{\mu\nu}$ and two-electron term $\mathbf{G}^x$. Note that, in addition to the conventional one-electron term ($H^x_{\mu\nu}$ in eq 4), there are two more terms on the right-hand side of eq 4, i.e., the ESP $V^x_{\mu\nu}$ arising from other fragment $K$ and the orbital projection operator, $P^i_{\mu\nu}$, which is made up of hybridized orbital $\varphi^h_i$ with $\gamma = 10^6$ hartree for all orbitals.[26]

Prior to the FMO calculations, the fragment borders are first defined appropriately.[45] The ESP is then obtained by iteratively solving the Fock equation of each monomer (eq 2 with $x = I$) until the energy of the monomers with the ESP of other fragments converge. Finally, every dimer (eq 2, $x = IJ$) is calculated once by using the ESP (eq 6 with $x = IJ$) obtained by the previous monomer FMO calculations. In the case of a polypeptide, a two-body expansion with a two residues per fragment division is a reasonable compromise between the attained accuracy and the computational cost incurred.[45] The FMO method has been widely used to study interactions such as dipole−dipole and $\pi-\pi$ interactions in biomolecular systems.[32,33,46]

**2.2. GIAO and CSGT Methods.** Chemical shifts can be evaluated theoretically by adding the contribution of the external magnetic field described by a vector potential to a Hamiltonian. Doing so, however, faces a new problem. Because the size of a basis set is finite, the calculated chemical shifts depend on the location of the gauge origin of the vector potential. To find a numerical solution to the nuclear magnetic shielding tensor which does not depend on the choice of the gauge origin, two methods have been established, i.e., GIAO[1,2] and CSGT.[3] With the GIAO method, the nuclear magnetic shielding tensor $\sigma_{\alpha\beta}$ is calculated from eq 9 as the mixed partial derivative of energy $E$ with respect to external magnetic field $\boldsymbol{B}$ and nuclear magnetic moment $\boldsymbol{\mu}$. The dependence of the gauge origin is mostly eliminated by using field-dependent atomic orbitals.

$$\sigma_{\alpha\beta} = \frac{\partial^2 E}{\partial \mu_\alpha \partial B_\beta} \tag{9}$$

In the CSGT method, $\boldsymbol{J}^{(1)}(\boldsymbol{r})$, the linear response of the current density induced by the external magnetic field $\boldsymbol{B}$ at the position $\boldsymbol{r}$ is determined using $\boldsymbol{r}$ as the gauge origin. The shielding at the nuclear position $\boldsymbol{r}_N$ is obtained by integrating the magnetic field (induction) generated by the induced current at $\boldsymbol{r}_N$ (eq 10).

$$\sigma_{\alpha\beta}(\boldsymbol{r}_{\mathrm{N}}) = -\frac{\mu_0}{4\pi} \int d^3\boldsymbol{r} \left[ \frac{\partial \boldsymbol{J}^{(1)}(\boldsymbol{r})}{\partial \boldsymbol{B}_\beta} \times \frac{\boldsymbol{r}_N - \boldsymbol{r}}{|\boldsymbol{r}_N - \boldsymbol{r}|^3} \right]_\alpha \tag{10}$$

**2.3. NMR Computational Models.** In a previous work, by combining the FMO method with either the GIAO or CSGT method, we developed two computational models for calculating the nuclear magnetic shielding tensor of large biomolecular systems, namely, model I and model II.[38] For both models, the Fock matrices of the monomers are calculated using the ESP of other fragments for close fragments within about 5 Å; otherwise, point charges are used. The two models differ in regard to the calculation of the Fock matrices of the dimers which are used to calculate chemical shifts. In model I, the dimer shielding tensors are calculated using a point charge ESP; that is, the point charge ESP from other fragments is constructed from the density matrix of monomers (Mulliken charges). Model II uses the ESP of other fragments without approximation. Fock matrices of dimers that are built on neighboring monomer pairs along the sequence of a given polypeptide are calculated by solving eq 2 for dimer $IJ$ (usually, $|I - J| = 1$). CPHF equations are then directly solved by using the Fock matrix to obtain the shielding tensor of the atoms in the dimer. GIAO is used in model I, whereas CSGT is used in both models I and II.[38]

Our previous tests[38] established that models I and II could reproduce the conventional quantum-chemical isotropic-shielding constants fairly well. Although model I could be applied to very large molecules, it created a larger error than model II because of the point-charge approximation. However, neither model could avoid larger errors in the case of the anisotropic shielding constants in comparison to those of conventional calculations. In fact, it is more difficult to evaluate anisotropic values than isotropic ones. Stikoff and

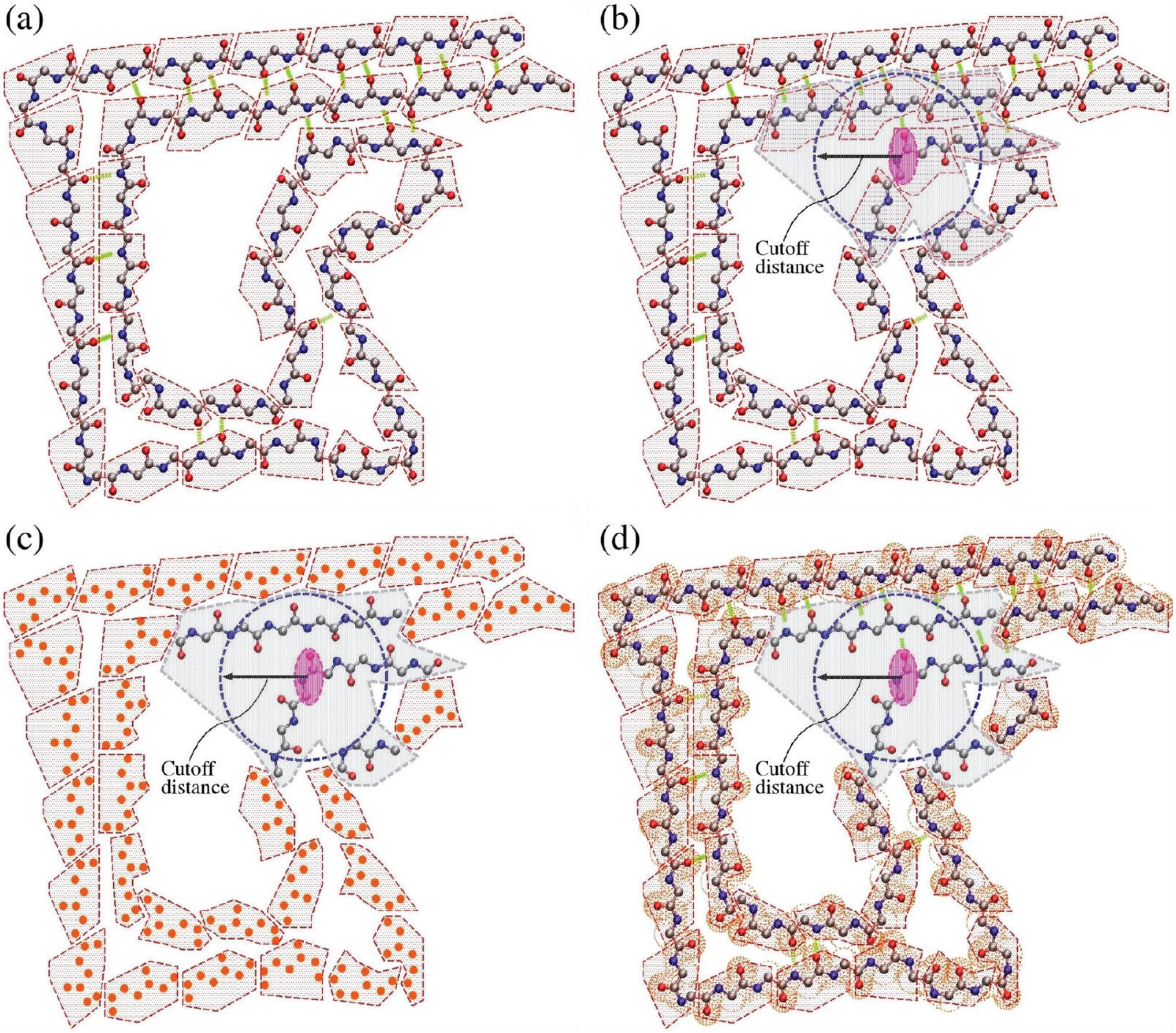


***Figure 1.*** Schematics of models I and II. The procedure of the calculation of models I and II consisting of three steps. Model I, step I (a): FMO1 calculations are performed (the backbone of a peptide is shown as an example). Each fragment consists of two residues, which is indicated by the gray area surrounded by a red dashed line. Step II (b): A cutoff distance (circled by the blue dashed line), measured from the center of mass of the target residue (indicated by pink oval), is used to build a merged fragment (enclosed by the silver dashed line). Step II (c): The merged fragment in the field of point charges given by FMO1 is calculated. Step III: NMR calculation of the merged fragment is performed using the density matrix from step II. Only chemical shifts of the target residue (pink oval) are retained. The NMR results for the entire molecule are obtained by repeating steps II and III for each residue. Model II, steps I and III are the same as for model I. Instead of step II (c) in model I, the merged fragment is calculated in the external Coulomb field determined by fragment densities from FMO1 (step II (d)), shown by schematic lobes.

Case[47] have already mentioned this difficulty; in particular, they stated that electrostatic effects (such as electrostatic polarization of bonds and noncovalent bond interactions) and magnetic susceptibilities (such as ring-current effects) of neighboring groups are the main reasons for the deviation in the nuclear magnetic shielding tensor.

In the following, we present a method by introducing a cutoff distance to determine the optimal fragment size in models I and II to achieve an accurate description of the electrostatic effects and magnetic susceptibilities. A schematic explanation of the new NMR computational method is shown in Figure 1. The point charges and the density matrix of each fragment (monomer) are first obtained by FMO-monomer (called FMO1) calculations (Figure 1a). A merged fragment is then constructed by assembling all the monomers within a cutoff distance (defined as $L_{\mathrm{cutoff}}$) from the center of mass (denoted by $\boldsymbol{R}_a$) of the residue under investigation (Figure 1b). Distance $R_{Ia}$ between given monomer $I$ and examined residue $a$ is chosen as the closest distance between atoms in $I$ and the center of mass of $a$ as follows:

$$R_{Ia} = \min_{i \in I} \{|\boldsymbol{R}_i - \boldsymbol{R}_a|\} \tag{11}$$

where $\boldsymbol{R}_i$ is the position of the $i$th atom in the $I$th monomer. A merged fragment (denoted as $Q(L_{\mathrm{cutoff}})$) including all the monomers within a given distance (eq 11) from residue $a$ is then created. In other words, if at least one of the atoms of the monomer is inside the area $R_{Ia} \leq L_{\mathrm{cutoff}}$, that monomer is assigned to the merged fragment, $Q(L_{\mathrm{cutoff}})$. The ESP,

$V_{\mu\nu}^{Q(L_{\text{cutoff}})}$, of fragment $Q(L_{\text{cutoff}})$ in models I (eq 12) and II (eq 13) is expressed as

$$V_{\mu\nu}^{Q(L_{\text{cutoff}})} = \sum_{K \notin Q(L_{\text{cutoff}})} \left\{ \sum_{A \in K} \left\langle \mu \left| -\frac{1}{4\pi\varepsilon_0} \frac{Z_A e^2}{|\boldsymbol{r} - \boldsymbol{R}_A|} \right| \nu \right\rangle + \sum_{A \in K} \left\langle \mu \left| \frac{1}{4\pi\varepsilon_0} \frac{Z_A^K e^2}{|\boldsymbol{r} - \boldsymbol{R}_A|} \right| \nu \right\rangle \right\} \tag{12}$$

$$V_{\mu\nu}^{Q(L_{\text{cutoff}})} = \sum_{K \notin Q(L_{\text{cutoff}})} \left\{ \sum_{A \in K} \left\langle \mu \left| -\frac{1}{4\pi\varepsilon_0} \frac{Z_A e^2}{|\boldsymbol{r} - \boldsymbol{R}_A|} \right| \nu \right\rangle + \sum_{\rho\sigma \in K} D_{\rho\sigma}^K (\mu\nu|\rho\sigma) \right\} \tag{13}$$

where $Z_A^K$ is the atomic population of atom $A$ on the $K$th monomer. (See also Figure 1c and d for models I and II, respectively.) Subsequently, the Fock matrix of $Q(L_{\text{cutoff}})$ is evaluated using the ESP in eqs 12 and 13.

$$\tilde{\boldsymbol{F}}^{Q(L_{\text{cutoff}})} = \tilde{\boldsymbol{H}}^{Q(L_{\text{cutoff}})} + \boldsymbol{G}^{Q(L_{\text{cutoff}})} \tag{14}$$

$$\tilde{H}_{\mu\nu}^{Q(L_{\text{cutoff}})} = H_{\mu\nu}^{Q(L_{\text{cutoff}})} + V_{\mu\nu}^{Q(L_{\text{cutoff}})} + \gamma \sum_{i \in Q(L_{\text{cutoff}})} P_{\mu\nu}^i \tag{15}$$

Finally, to obtain magnetic shielding tensors, the Fock matrix is used to solve the CPHF equations. Here, the new cutoff-distance-based method is denoted as “FMO1(merged)” (where “merged” stands for the use of a merged fragment), whereas the dimer-based method in the previous work[38] is denoted as “FMO1(dimer).”

## 3. Computational Details

**3.1. Structural Modeling of Polypeptides.** The α-helix and β-sheet, which are the two basic secondary structural elements in proteins, were selected to evaluate the performance of the FMO1(merged) method. An α-helix peptide was chosen from one of the α-helices (residues 241 to 272 of chain A) in bovine rhodopsin (Protein Data Bank code: 1HZX[48]). This α-helix structure is stabilized by the hydrogen bonds in its main chain (Figure 2). The β-sheet peptide was extracted from residues 198 to 229 in the green fluorescent protein (PDB code: 1Q4B[49]). The β-sheet structure is formed from two β-strands, which are connected by a loop structure. As shown in Figure 2, the β-sheet has a more extended structure than the α-helix.

Hydrogen atoms were added to both the α-helix and β-sheet in a pH = 7 environment by using the leap module in Amber 8.[50] These two structures were minimized with the steepest decent and conjugated gradient methods for 2500 steps by using the Amber99 force field.[51] After that, MOZYME[52] with the AM1[53] Hamiltonian in MOPAC2007[54] was used to optimize these two peptides with a threshold value of 10 Å in the NDDO approximation. The same β-sheet structure was employed in ref 40, where the accuracy of the chemical shifts calculated by FMO1(dimer)/model I and FMO1(dimer)/model II was assessed.

**3.2. FMO1(merged) Calculations.** The computational process for FMO1(merged)/model I and FMO1(merged)/model II consists of three steps. It is explained in regard to model I as follows. Step I: The polypeptide is divided into several fragments (monomers) by assigning two adjacent residues to one fragment. The electron density of all monomers is computed by performing an FMO calculation at the monomer level (FMO1 calculations). The point charges of all atoms are estimated by using the Mulliken approximation.[55] Step II: For each residue, $Q(L_{\text{cutoff}})$ is determined by using $L_{\text{cutoff}}$. Using the ESP of eq 12, the wave function of $Q(L_{\text{cutoff}})$ is obtained. Step III: Using the wave function obtained in step II, CPHF calculations are performed with GIAO or CSGT methods to determine the NMR magnetic shielding tensors of $Q(L_{\text{cutoff}})$.

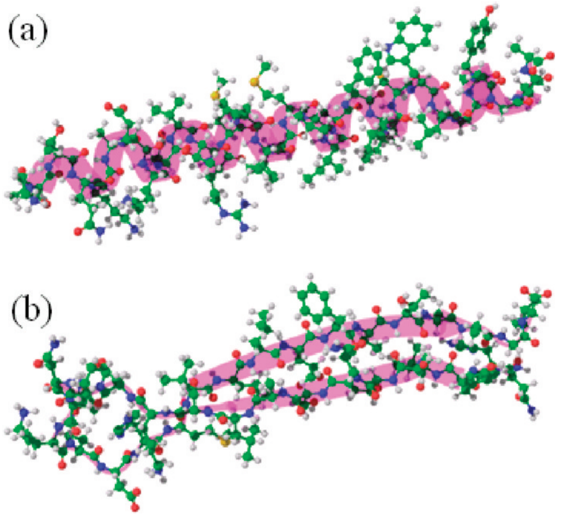


*Figure 2.* (a) α-Helix and (b) β-sheet polypeptide structures used for calculating NMR chemical shifts in this study. Peptide sequence of the α-helix is Ala-Thr-Thr-Gln-Lys-Ala-Glu-Lys-Glu-Val-Thr-Arg-Met-Val-Ile-Ile-Met-Val-Ile-Ala-Phe-Leu-Ile-Cys-Trp-Leu-Pro-Tyr-Ala-Gly-Val-Ala. Peptide sequence of the β-sheet is Asn-His-Tyr-Leu-Ser-Thr-Gln-Ser-Ala-Leu-Ser-Lys-Asp-Pro-Asn-Glu-Lys-Arg-Asp-His-Met-Val-Leu-Leu-Glu-Phe-Val-Thr-Ala-Ala-Gly-Ile.

In regard to model II, steps I and III are the same as for model I. The difference from model I appears in step II. According to eq 13, the monomer density matrices are used to evaluate the ESP outside $L_{\text{cutoff}}$. The CSGT method is used for model II (see section 2.3).

To obtain the NMR shielding tensors of the polypeptide, steps II and III are performed for all the residues. The isotropic and anisotropic shielding constants are evaluated from the shielding tensors. The anisotropic shielding constant is defined as

$$\Delta\sigma = \sigma_3 - (\sigma_1 + \sigma_2)/2 \tag{16}$$

where $\sigma_1$, $\sigma_2$, and $\sigma_3$ are the eigenvalues of nuclear magnetic shielding tensor $\sigma$, and $\sigma_3$ is the largest of these.

The dependence of shielding constants on $L_{\text{cutoff}}$ (6, 8, and 10 Å) used in FMO1(merged)/model I and FMO1(merged)/model II is investigated. For comparison, the shielding constants calculated without the $V_{\mu\nu}^{Q(L_{\text{cutoff}})}$ term, that is, the shielding constants of the merged molecule surrounded by neither point changes nor the ESP of other fragments, are shown. Further-

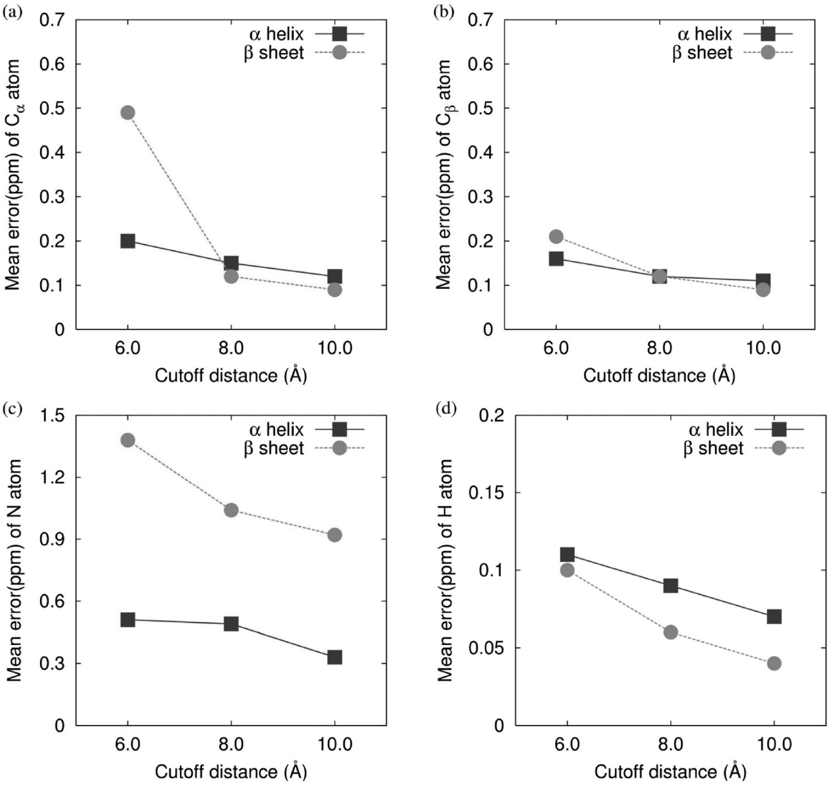


***Figure 3.*** Mean error of isotropic shielding constants as a function of cutoff distances obtained by FMO1(merged)/model I with CSGT calculations (6-31G(d)) in the case of the α-helix (solid line with squares) and the $\beta$-sheet (dashed line with circles). (a)$^{13}C_{\alpha}$, (b) $^{13}C_{\beta}$, (c) $^{15}N$, and (d) $^{1}H$ atoms.

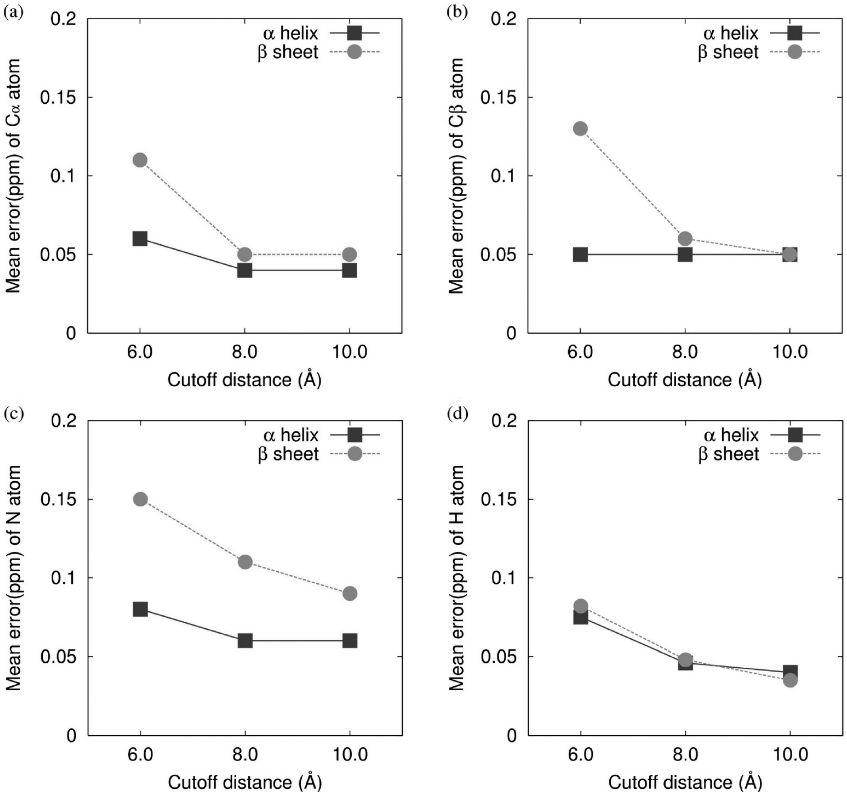


***Figure 4.*** Mean error of isotropic shielding constants as a function of cutoff distances obtained by FMO1(merged)/model II with CSGT calculations (6-31G(d)) in the case of the α-helix (solid line with squares) and the $\beta$-sheet (dashed line with circles). (a)$^{13}C_{\alpha}$, (b) $^{13}C_{\beta}$, (c) $^{15}N$, and (d) $^{1}H$ atoms.

more, the effect of basis sets (6-31G(d) and 6-311G(d,p)) on the shielding constants was tested. The FMO and NMR (GIAO and CSGT) calculations were performed using GAMESS[56] and Gaussian 03,[57] respectively. In the following, the errors of

***Table 1.*** Quality of Absolute Isotropic NMR Shielding Constants (in ppm) of $^{13}C_\alpha$, $^{13}C_\beta$, $^{15}N$, and $^{1}H$ Atoms in the α-Helix and β-Sheet Calculated with FMO1(merged)/model I by Using CSGT and the 6-31G(d) Basis Set

| atoms | $^{13}C_\alpha$ | | | $^{13}C_\beta$ | | | $^{15}N$ | | | $^{1}H$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cutoff (Å) | 6.0 | 8.0 | 10.0 | 6.0 | 8.0 | 10.0 | 6.0 | 8.0 | 10.0 | 6.0 | 8.0 | 10.0 |
| α-Helix in Bovine Rhodopsin 1HZX (residues 241−272) | | | | | | | | | | | | |
| max error | 0.58 | 0.33 | 0.26 | 0.41 | 0.31 | 0.31 | 2.19 | 2.19 | 1.16 | 0.26 | 0.20 | 0.17 |
| mean error | 0.20 | 0.15 | 0.12 | 0.16 | 0.12 | 0.11 | 0.51 | 0.49 | 0.33 | 0.11 | 0.09 | 0.07 |
| standard deviation | 0.14 | 0.09 | 0.08 | 0.11 | 0.08 | 0.08 | 0.40 | 0.39 | 0.26 | 0.07 | 0.06 | 0.04 |
| β-Sheet in Green Fluorescent Protein 1Q4B (residues 198−229) | | | | | | | | | | | | |
| max error | 1.74 | 0.26 | 0.26 | 0.92 | 0.49 | 0.41 | 3.56 | 3.01 | 2.21 | 0.31 | 0.14 | 0.11 |
| mean error | 0.49 | 0.12 | 0.09 | 0.21 | 0.12 | 0.09 | 1.38 | 1.04 | 0.92 | 0.10 | 0.06 | 0.04 |
| standard deviation | 0.54 | 0.08 | 0.07 | 0.20 | 0.11 | 0.09 | 0.96 | 0.78 | 0.63 | 0.08 | 0.03 | 0.03 |

***Table 2.*** Quality of Absolute Anisotropic NMR Shielding Constants (in ppm) of $^{13}C_\alpha$, $^{13}C_\beta$, $^{15}N$, and $^{1}H$ Atoms in the α-Helix and β-Sheet Calculated with FMO1(merged)/model I by Using CSGT and the 6-31G(d) Basis Set

| atoms | $^{13}C_\alpha$ | | | $^{13}C_\beta$ | | | $^{15}N$ | | | $^{1}H$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cutoff (Å) | 6.0 | 8.0 | 10.0 | 6.0 | 8.0 | 10.0 | 6.0 | 8.0 | 10.0 | 6.0 | 8.0 | 10.0 |
| α-Helix in Bovine Rhodopsin 1HZX (residues 241−272) | | | | | | | | | | | | |
| max error | 1.42 | 1.42 | 0.77 | 0.99 | 0.62 | 0.62 | 1.52 | 1.52 | 1.06 | 1.96 | 1.37 | 0.70 |
| mean error | 0.67 | 0.48 | 0.32 | 0.29 | 0.21 | 0.17 | 0.59 | 0.47 | 0.36 | 1.10 | 0.68 | 0.43 |
| standard deviation | 0.35 | 0.32 | 0.22 | 0.25 | 0.19 | 0.14 | 0.41 | 0.35 | 0.25 | 0.54 | 0.31 | 0.17 |
| β-Sheet in Green Fluorescent Protein 1Q4B (residues 198−229) | | | | | | | | | | | | |
| max error | 2.40 | 1.54 | 1.13 | 1.92 | 0.72 | 0.59 | 2.98 | 2.07 | 1.60 | 2.56 | 1.03 | 0.85 |
| mean error | 1.02 | 0.65 | 0.45 | 0.54 | 0.25 | 0.17 | 0.95 | 0.47 | 0.37 | 0.64 | 0.42 | 0.27 |
| standard deviation | 0.62 | 0.37 | 0.28 | 0.46 | 0.19 | 0.15 | 0.77 | 0.44 | 0.32 | 0.56 | 0.28 | 0.20 |

***Table 3.*** Quality of Absolute Isotropic NMR Shielding Constants (in ppm) of $^{13}C_\alpha$, $^{13}C_\beta$, $^{15}N$, and $^{1}H$ Atoms in the α-Helix and β-Sheet Calculated with FMO1(merged)/model II by Using CSGT and the 6-31G(d) Basis Set

| atoms | $^{13}C_\alpha$ | | | $^{13}C_\beta$ | | | $^{15}N$ | | | $^{1}H$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cutoff (Å) | 6.0 | 8.0 | 10.0 | 6.0 | 8.0 | 10.0 | 6.0 | 8.0 | 10.0 | 6.0 | 8.0 | 10.0 |
| α-Helix in Bovine Rhodopsin 1HZX (residues 241−272) | | | | | | | | | | | | |
| max error | 0.18 | 0.12 | 0.12 | 0.13 | 0.13 | 0.10 | 0.24 | 0.20 | 0.20 | 0.15 | 0.09 | 0.08 |
| mean error | 0.06 | 0.04 | 0.04 | 0.05 | 0.05 | 0.05 | 0.08 | 0.06 | 0.06 | 0.08 | 0.05 | 0.04 |
| standard deviation | 0.04 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.06 | 0.05 | 0.05 | 0.04 | 0.03 | 0.02 |
| β-Sheet in Green Fluorescent Protein 1Q4B (residues 198−229) | | | | | | | | | | | | |
| max error | 0.26 | 0.15 | 0.15 | 0.48 | 0.16 | 0.15 | 0.68 | 0.24 | 0.24 | 0.29 | 0.11 | 0.09 |
| mean error | 0.11 | 0.05 | 0.05 | 0.13 | 0.06 | 0.05 | 0.15 | 0.11 | 0.09 | 0.08 | 0.05 | 0.04 |
| standard deviation | 0.07 | 0.04 | 0.03 | 0.13 | 0.04 | 0.04 | 0.15 | 0.07 | 0.07 | 0.07 | 0.03 | 0.02 |

***Table 4.*** Quality of Absolute Anisotropic NMR Shielding Constants (in ppm) of $^{13}C_\alpha$, $^{13}C_\beta$, $^{15}N$, and $^{1}H$ Atoms in the α-Helix and β-Sheet Calculated with FMO1(merged)/model II by Using CSGT and the 6-31G(d) Basis Set

| atoms | $^{13}C_\alpha$ | | | $^{13}C_\beta$ | | | $^{15}N$ | | | $^{1}H$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cutoff (Å) | 6.0 | 8.0 | 10.0 | 6.0 | 8.0 | 10.0 | 6.0 | 8.0 | 10.0 | 6.0 | 8.0 | 10.0 |
| α-Helix in Bovine Rhodopsin 1HZX (residues 241−272) | | | | | | | | | | | | |
| max error | 0.77 | 0.77 | 0.48 | 0.77 | 0.47 | 0.47 | 0.90 | 0.77 | 0.54 | 1.86 | 1.36 | 0.71 |
| mean error | 0.44 | 0.32 | 0.21 | 0.26 | 0.19 | 0.13 | 0.42 | 0.32 | 0.20 | 1.08 | 0.67 | 0.42 |
| standard deviation | 0.20 | 0.15 | 0.11 | 0.18 | 0.14 | 0.11 | 0.25 | 0.21 | 0.15 | 0.53 | 0.30 | 0.18 |
| β-Sheet in Green Fluorescent Protein 1Q4B (residues 198−229) | | | | | | | | | | | | |
| max error | 2.44 | 1.07 | 0.85 | 1.27 | 0.69 | 0.53 | 1.71 | 0.69 | 0.61 | 2.21 | 0.93 | 0.73 |
| mean error | 1.05 | 0.52 | 0.36 | 0.45 | 0.25 | 0.14 | 0.75 | 0.37 | 0.28 | 0.57 | 0.37 | 0.24 |
| standard deviation | 0.60 | 0.26 | 0.21 | 0.33 | 0.16 | 0.13 | 0.47 | 0.19 | 0.15 | 0.49 | 0.24 | 0.18 |

shielding constants are reported as deviations compared to the shielding constants by conventional *ab initio* methods.

## 4. Results and Discussion

**4.1. Accuracy Comparison of FMO1(merged) and FMO1(dimer).** The errors in shielding constants using FMO1(merged) are much smaller than those obtained with the previously developed FMO1(dimer). These errors are summarized in Tables 1−4. In the case of FMO1(merged)/model I, both the isotropic and anisotropic shielding constants of carbon atoms ($^{13}C_\alpha$ and $^{13}C_\beta$) agree well with those determined by conventional *ab initio* NMR calculations. Although FMO1(dimer)/model II gives isotropic shielding constants with a larger maximum error (less than 3.88 ppm) and mean error (less than 0.89 ppm),[40] FMO1(merged)/model II with CSGT ($L_{\mathrm{cutoff}} = 10$ Å) reduces the maximum

***Table 5.*** Quality of Absolute Isotropic NMR Shielding Constants (in ppm) of $^{13}C_\alpha$, $^{13}C_\beta$, $^{15}N$, and $^{1}H$ Atoms in the α-Helix and $\beta$-Sheet Calculated with FMO1(merged) without the $V_{\mu\nu}^{Q(L_{cutoff})}$ Term by Using CSGT and the 6-31G(d) Basis Set

| atoms | $^{13}C_\alpha$ | | | $^{13}C_\beta$ | | | $^{15}N$ | | | $^{1}H$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cutoff (Å) | 6.0 | 8.0 | 10.0 | 6.0 | 8.0 | 10.0 | 6.0 | 8.0 | 10.0 | 6.0 | 8.0 | 10.0 |
| α-Helix in Bovine Rhodopsin 1HZX (residues 241−272) | | | | | | | | | | | | |
| max error | 0.51 | 0.31 | 0.31 | 0.40 | 0.40 | 0.28 | 2.13 | 1.18 | 1.01 | 0.48 | 0.19 | 0.19 |
| mean error | 0.16 | 0.10 | 0.08 | 0.12 | 0.10 | 0.08 | 0.64 | 0.41 | 0.26 | 0.16 | 0.07 | 0.07 |
| standard deviation | 0.11 | 0.10 | 0.08 | 0.11 | 0.10 | 0.08 | 0.48 | 0.34 | 0.22 | 0.11 | 0.05 | 0.04 |
| $\beta$-Sheet in Green Fluorescent Protein 1Q4B (residues 198−229) | | | | | | | | | | | | |
| max error | 3.00 | 0.64 | 0.42 | 1.24 | 0.55 | 0.41 | 7.52 | 4.46 | 2.85 | 0.32 | 0.20 | 0.18 |
| mean error | 0.64 | 0.11 | 0.07 | 0.22 | 0.10 | 0.08 | 1.53 | 1.07 | 0.94 | 0.10 | 0.04 | 0.03 |
| standard deviation | 0.74 | 0.12 | 0.06 | 0.27 | 0.12 | 0.10 | 1.81 | 1.02 | 0.59 | 0.08 | 0.04 | 0.04 |

***Table 6.*** Quality of Absolute Anisotropic NMR Shielding Constants (in ppm) of $^{13}C_\alpha$, $^{13}C_\beta$, $^{15}N$, and $^{1}H$ Atoms in the α-Helix and $\beta$-Sheet Calculated with FMO1(merged) without the $V_{\mu\nu}^{Q(L_{cutoff})}$ Term by Using CSGT and the 6-31G(d) Basis Set

| atoms | $^{13}C_\alpha$ | | | $^{13}C_\beta$ | | | $^{15}N$ | | | $^{1}H$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| cutoff (Å) | 6.0 | 8.0 | 10.0 | 6.0 | 8.0 | 10.0 | 6.0 | 8.0 | 10.0 | 6.0 | 8.0 | 10.0 |
| α-Helix in Bovine Rhodopsin 1HZX (residues 241−272) | | | | | | | | | | | | |
| max error | 1.33 | 1.17 | 0.71 | 1.00 | 0.60 | 0.58 | 1.00 | 0.75 | 0.78 | 2.04 | 1.40 | 0.79 |
| mean error | 0.64 | 0.37 | 0.31 | 0.31 | 0.21 | 0.16 | 0.40 | 0.31 | 0.30 | 1.16 | 0.70 | 0.44 |
| standard deviation | 0.38 | 0.25 | 0.26 | 0.26 | 0.20 | 0.14 | 0.26 | 0.20 | 0.22 | 0.59 | 0.31 | 0.19 |
| $\beta$-Sheet in Green Fluorescent Protein 1Q4B (residues 198−229) | | | | | | | | | | | | |
| max error | 2.24 | 1.24 | 1.04 | 3.42 | 0.81 | 0.68 | 3.13 | 2.27 | 1.74 | 2.67 | 1.02 | 0.86 |
| mean error | 1.00 | 0.55 | 0.40 | 0.76 | 0.34 | 0.20 | 0.92 | 0.48 | 0.33 | 0.65 | 0.41 | 0.27 |
| standard deviation | 0.58 | 0.33 | 0.28 | 0.67 | 0.21 | 0.16 | 0.76 | 0.41 | 0.31 | 0.59 | 0.28 | 0.22 |

and absolute mean errors of isotropic shielding constants of carbon atoms in the $\beta$-sheet, i.e., to less than 0.15 and 0.05 ppm, respectively, with the 6-31G(d) basis set (Table 3). And it obtains a maximum error of less than 0.85 ppm and a mean error of less than 0.36 ppm with this basis set for the anisotropic shielding constants of carbon atoms (Table 4). These errors for the anisotropic shielding constants are, again, much smaller than those with FMO1(dimer)/model II. Similarly to the case of model II, the errors for FMO1(merged)/model I (Tables 1 and 2) are much smaller than those for the FMO1(dimer)/model I calculations on $^{13}C$ atoms.

The errors of the shielding constants for $^{15}N$ and $^{1}H$ atoms by using FMO1(merged) are remarkably reduced. For example, the maximum errors in isotropic shielding constants (determined by using FMO1 (merged)/with model II (CSGT) ($L_{cutoff} = 10$ Å) with the 6-31G(d) basis set) of the $\beta$-sheet are 0.24 and 0.09 ppm for $^{15}N$ and $^{1}H$, respectively (Table 3). In contrast, the previously reported[40] maximum errors in isotropic shielding constants (determined by using FMO1(dimer)/model II (CSGT) with the same basis sets) were 7.51 and 0.96 ppm for $^{15}N$ and $^{1}H$, respectively. Because the shielding constants of these two atoms are more sensitive to the surrounding chemical environment than that of $^{13}C$ atoms, such a large error must have been produced by the inaccuracy of the FMO1(dimer) models in reproducing the surrounding chemical environment.

These results indicate that FMO1(merged) provides a much more accurate electrostatic description and magnetic susceptibilities around the atom of interest than those provided by FMO1(dimer). This improved accuracy is discussed further in sections 4.5 to 4.8.

**4.2. Effect of Cutoff Distance.** The effect of cutoff distance on the accuracy of shielding constants is investigated in the following. This investigation focused on two questions: First, how large should the cutoff distance be to accurately reproduce conventional *ab initio* results? Second, does the effect of cutoff distance depend on the choice of polypeptide structures?

The shielding constants of the α-helix structure were calculated by using FMO1(merged)/model I with CSGT and the 6-31G(d) basis set. The calculation results are compared with the results obtained with the conventional *ab initio* method in Tables 1 and 2. For the most environmentally sensitive anisotropic shielding constants of the α-helix, the larger the $L_{cutoff}$ (from 6 to 10 Å), the smaller the mean error. A similar trend is also found in regard to the isotropic shielding constants.

The shielding constants of the $\beta$-sheet were also calculated using FMO1(merged)/model I. The errors of anisotropic shielding constants for $^{1}H$ atoms in the $\beta$-sheet are smaller than those for $^{1}H$ atoms in the α-helix, while those for $^{13}C$ and $^{15}N$ atoms in the $\beta$-sheet are slightly larger. A significantly larger decrease in mean error is found in the case of the $\beta$-sheet than in the case of the α-helix when $L_{cutoff}$ increased from 6 to 8 Å (Figure 3). In contrast, when the cutoff distance increases from 8 to 10 Å, the decrease in the mean error in the isotropic values for the α-helix and $\beta$-sheet is almost the same. A similar decrease in error is also found for the shielding constants of $^{13}C_\alpha$, $^{13}C_\beta$, and $^{15}N$ atoms calculated with FMO1(merged)/model II (Figure 4). For $^{1}H$ atoms, the decrease in error is almost the same from 6 to 10 Å.

We suggest that the cutoff-distance dependency of the chemical shifts is due to the secondary structures. The structure of the $\beta$-sheet is more extended than that of the α-helix. Therefore, in regard to the total number of residues sequentially along the polypeptide (counted from the residue where the chemical shifts are calculated to the residue at

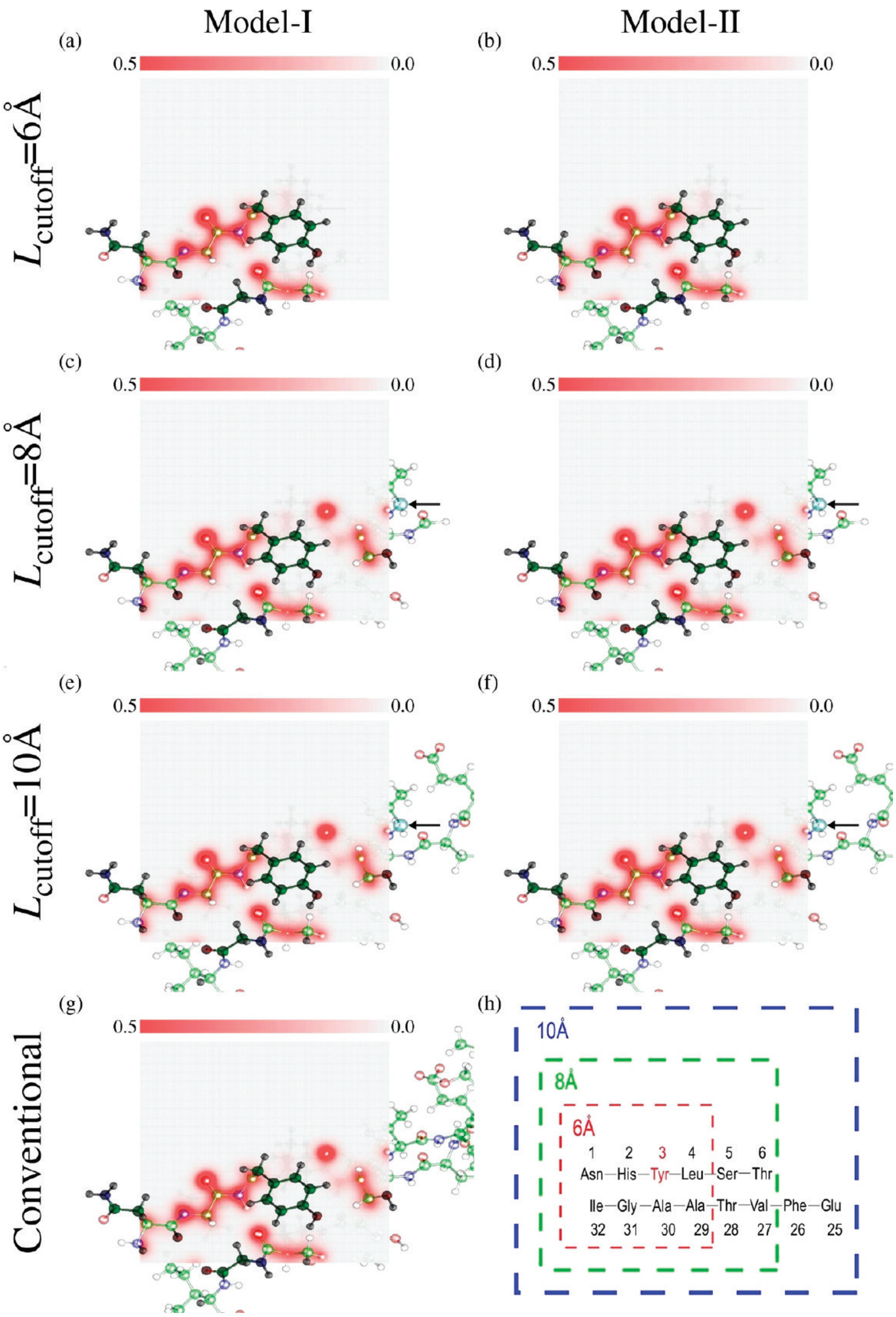


***Figure 5.*** Charge density on the surface specified by three atoms (C, O, and N) of the peptide bond between His2 and Tyr3 in the $\beta$-sheet. The magnitude of the charge density is expressed in proportion to the intensity of the color (red), which is indicated as the gradation of the color at the top of each panel. The unit is in $e/(a_0)^3$, where $a_0$ is the Bohr radius and the area is 15.7 × 15.7 Å$^2$. FMO1(merged)/model I results obtained using CSGT (6-31G(d)) with $L_{cutoff}$ values of (a) 6 Å, (c) 8 Å, and (e) 10 Å and FMO1(merged)/model II results obtained using CSGT (6-31G(d)) with $L_{cutoff}$ values of (b) 6 Å, (d) 8 Å, and (f) 10 Å. (g) Results of conventional *ab initio* calculation using CSGT (6-31G(d)). One of the border atoms at the fragment boundary is represented as a light-green transparent sphere and indicated by an arrow for each panel. (h) Residues within $L_{cutoff}$ from the center of mass of Try3 are shown.

the FMO boundary), the number is larger in the α-helix than in the $\beta$-sheet. In other words, if the same cutoff distance is used for the α-helix and the $\beta$-sheet, the effect of the FMO boundary is larger for the $\beta$-sheet than for the α-helix. The effect of the FMO boundary, however, is sufficiently small for the cutoff distance of 8 Å; thus, the magnitude of errors is similar for the α-helix and $\beta$-sheet for $L_{cutoff} \geq 8$ Å.

**4.3. Effect of Basis Sets.** To assess the dependence of the cutoff distance on the type of basis set, basis sets 6-31G(d) and 6-311G(d,p) were used. (As is well-known, a large basis set is needed to obtain accurate chemical-shift values.) The shielding constants of the α-helix and $\beta$-sheet polypeptides, which were calculated with the 6-311G(d,p) basis set and FMO1(merged) and the conventional method, are compared (Tables S1−S4, Supporting Information). The magnitude of errors for $L_{cutoff} = 8$ Å is converged to small values irrespective of the size of the basis set. The magnitude of errors for $L_{cutoff} = 6$ Å in

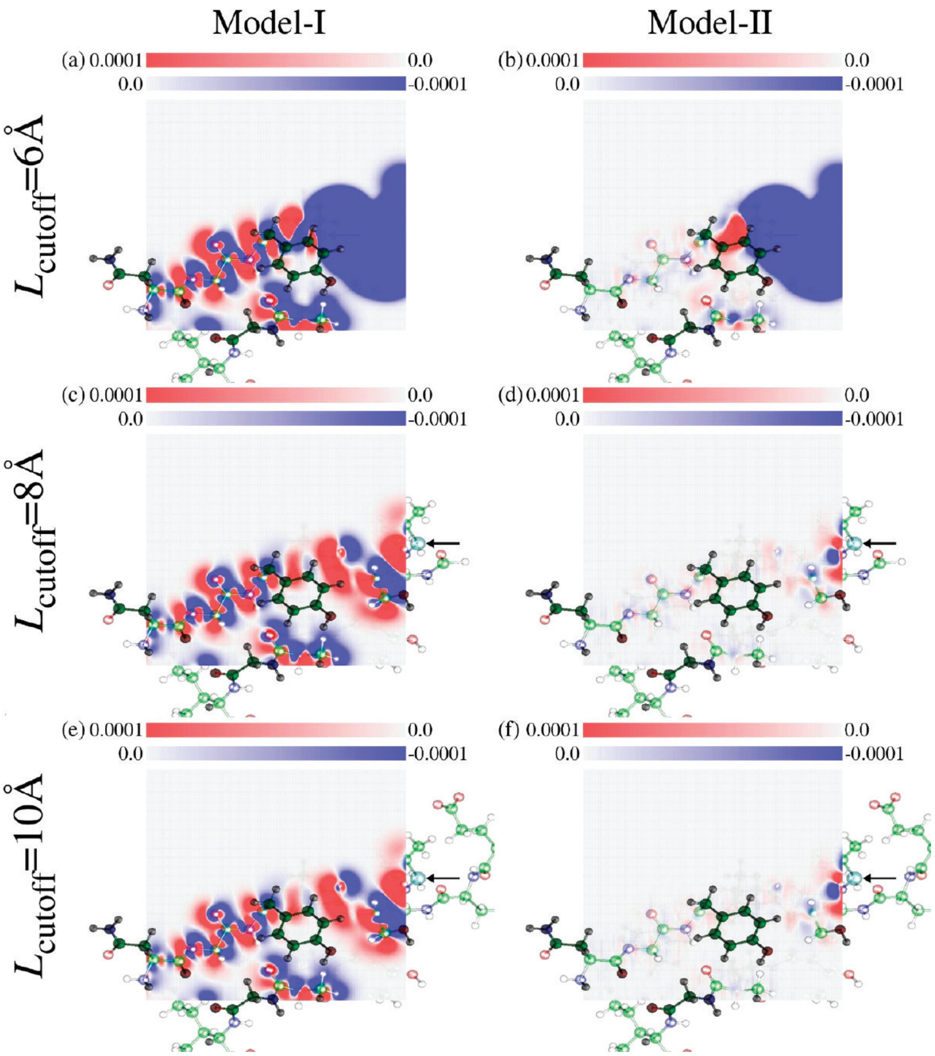


***Figure 6.*** Differences between charge densities determined by FMO1(merged) and conventional *ab initio* calculations (6-31G(d)) given in Figure 5. The differences are calculated as $\Delta\rho = \rho_{\mathrm{FMO1(merged)}} - \rho_{\mathrm{conventional}}$. The magnitude of the difference in charge density is expressed in proportion to the intensity of the colors (red for positive values and blue for negative valuse), which are indicated as the gradation of the colors at the top of each panel. The unit is $e/(a_0)^3$, where $a_0$ is the Bohr radius. Area is 15.7 × 15.7 Å$^2$. The FMO1(merged) calculation results are obtained using model I with CSGT with $L_{\mathrm{cutoff}}$ values of (a) 6 Å, (c) 8 Å, and (e) 10 Å and model II with CSGT with $L_{\mathrm{cutoff}}$ values of (b) 6 Å, (d) 8 Å, and (f) 10 Å. One of the border atoms at the fragment boundary is represented as a light-green transparent sphere and indicated by an arrow for each panel.

the case of basis set 6-311G(d,p) is larger than that in the case of basis set 6-31G(d). Accordingly, $L_{\mathrm{cutoff}} \geq 8$ Å is recommended as the appropriate value.

**4.4. Comparison of Accuracy of Models I and II.** Model I produces a larger error than model II due to the point-charge approximation. For example, the maximum error of $^{15}$N isotropic shielding constants relative to the conventional *ab initio* values is 1.16−2.19 ppm with a mean error of 0.33−0.51 ppm (Table 1) in the case of the FMO1(merged)/model I calculation of the α-helix with the 6-31G(d) basis set. The maximum absolute error of $^{15}$N isotropic shielding constants is about 300%−400% larger than the other errors of the $^{15}$N isotropic shielding constants for the α-helix. In contrast, FMO1(merged)/model II gives shielding constants that are in excellent agreement with the conventional *ab initio* results (maximum errors, 0.20−0.24 ppm; mean errors, 0.06−0.08 ppm for $^{15}$N isotropic shielding constants of the α-helix (Table 3)).

The accuracy of shielding constants calculated by FMO1(merged)/model I and FMO1(merged)/model II with different values of $L_{\mathrm{cutoff}}$ for basis set 6-31G(d) was analyzed. Model II provides much more accurate results than model I especially for isotropic values. For example, the mean error in the isotropic shielding constants of $^{15}$N in the α-helix calculated with model II ($\bar{\varepsilon}_{\mathrm{Model-II}}$) for $L_{\mathrm{cutoff}} = 6$ Å is significantly reduced compared to that calculated with model I ($\bar{\varepsilon}_{\mathrm{Model-I}}$) by a factor of $\bar{\varepsilon}_{\mathrm{Model-I}}/\bar{\varepsilon}_{\mathrm{Model-II}} = 6.38$ (Tables 1 and 3). The larger error with model I is due to the use of Mulliken charges. This causes somewhat larger errors, especially for $^{15}$N and $^{1}$H atoms. The ineffectiveness of the use of Mulliken charges is clearly indicated by comparing the errors given by FMO1(merged)/model I (Tables 1 and 2) and FMO1(merged) without the $V_{\mu\nu}^{Q(L_{\mathrm{cutoff}})}$ term (Tables 5 and 6). The errors with FMO1(merged)/model I are similar to those with FMO1(merged) without the $V_{\mu\nu}^{Q(L_{\mathrm{cutoff}})}$ term. In contrast, the errors in isotropic shielding constants calculated

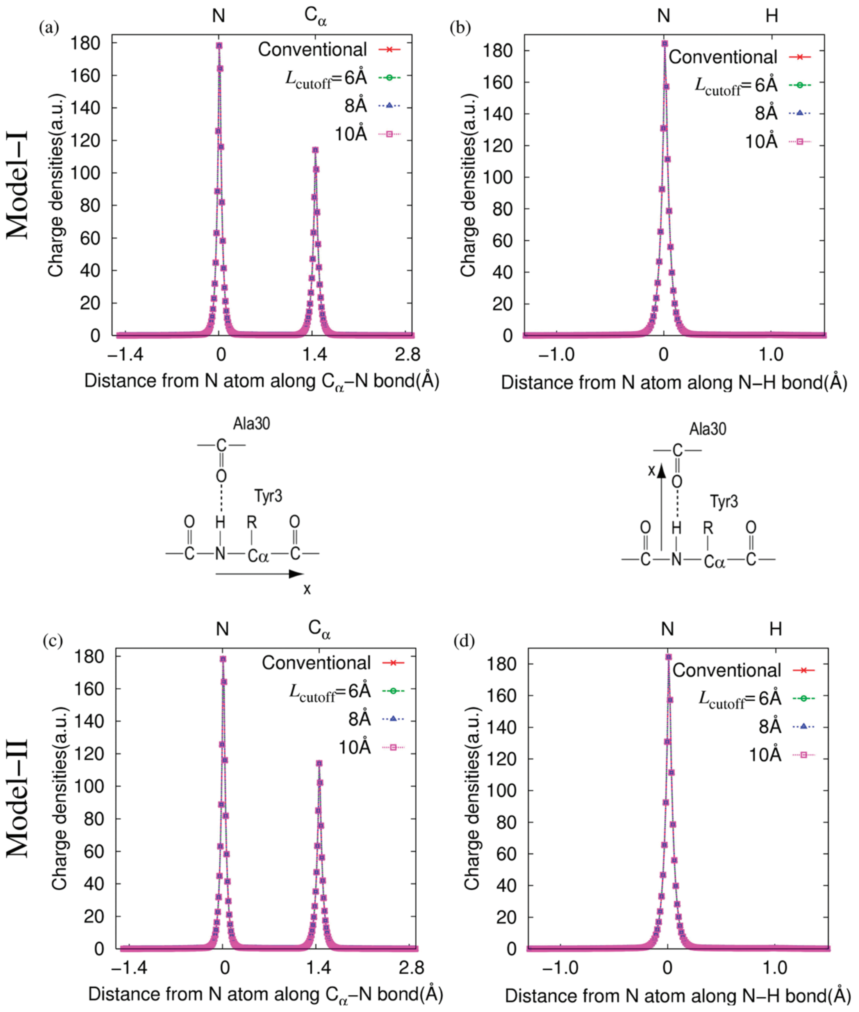


*Figure 7.* Charge densities plotted along $C_\alpha$−N and N−H bonds. Results calculated with the conventional *ab initio* method and FMO1(merged)/model I with CSGT(6-31G(d)) (a) along the $C_\alpha$−N bond and (b) along the N−H bond. Results calculated with the conventional *ab initio* method and the FMO1(merged)/model II with CSGT (6-31G(d)) (c) along the $C_\alpha$−N bond and (d) along the N−H bond. Because the curves are hard to distinguish at the scale shown, the differences are shown in Figure 8.

by FMO1(dimer)/model I are smaller than those calculated by FMO1(dimer) without the $V_{\mu\nu}^{Q(L_{cutoff})}$ term. This is due to the fact that the effect of the ESP created by the surrounding fragments is large in the FMO1(dimer) calculations but is relatively smaller in FMO1(merged) calculations at least for $L_{cutoff} \geq 6$ Å. Thus, to improve the results calculated by FMO1(merged) for large cutoff distances, the ESP due to the surrounding fragments has to be reproduced with a high degree of precision.

**4.5. Charge-Density Distribution.** To examine the electronic structures obtained by models I and II, the distribution of the charge density in the critical area was analyzed, i.e., around the Tyr3 residue in the $\beta$-sheet, in which the maximum error in the isotropic values of $^{15}$N was found (see Figure 5 and Table S9, Supporting Information).

The charge distribution calculated with FMO1(merged) on the surface specified by three atoms (C, O, and N) of the peptide bond between His2 and Tyr3 in the $\beta$-sheet is given in Figure 5a−f. The magnitude of the charge density is expressed in proportion to the intensity of the color (red). The charge distributions are indistinguishable from those determined by the conventional *ab initio* calculations (Figure 5g), except for the charge distributions in the areas on the right in Figure 5a and b ($L_{cutoff} = 6$ Å).

To clarify the difference between the charge distributions calculated in the case of three $L_{cutoff}$ values in both models, the difference in charge densities ($\Delta\rho = \rho_{FMO1(merged)} - \rho_{conventional}$) is given in Figure 6. As also shown in Figure 5, the change in the distribution of charge density is due to the change in $L_{cutoff}$ from 6 to 8 Å (Figures 6a and b for an $L_{cutoff}$ of 6 Å and c and d for an $L_{cutoff}$ of 8 Å), whereas the change is small when $L_{cutoff}$ is increased from 8 to 10 Å (Figures 6e and f). Although the difference between models I and II is not clear in Figure 5, it is obvious in Figure 6, which plots error $\Delta\rho$. There is a large deviation in the charge density around the heavy atoms. The deviation in the case of model I (Figure 6a, c, and e) is more complicated than that in the case of model II (Figures 6b, d, and f).

The charge distributions along the $C_\alpha$−N and N−H bonds of Tyr3 in the $\beta$-sheet were examined in more detail (Figure 7). The charge densities calculated with FMO1(merged)/model I and FMO1(merged)/model II are almost identical with those calculated with the conventional *ab initio* method for all cutoff distances. Moreover, the differences in charge

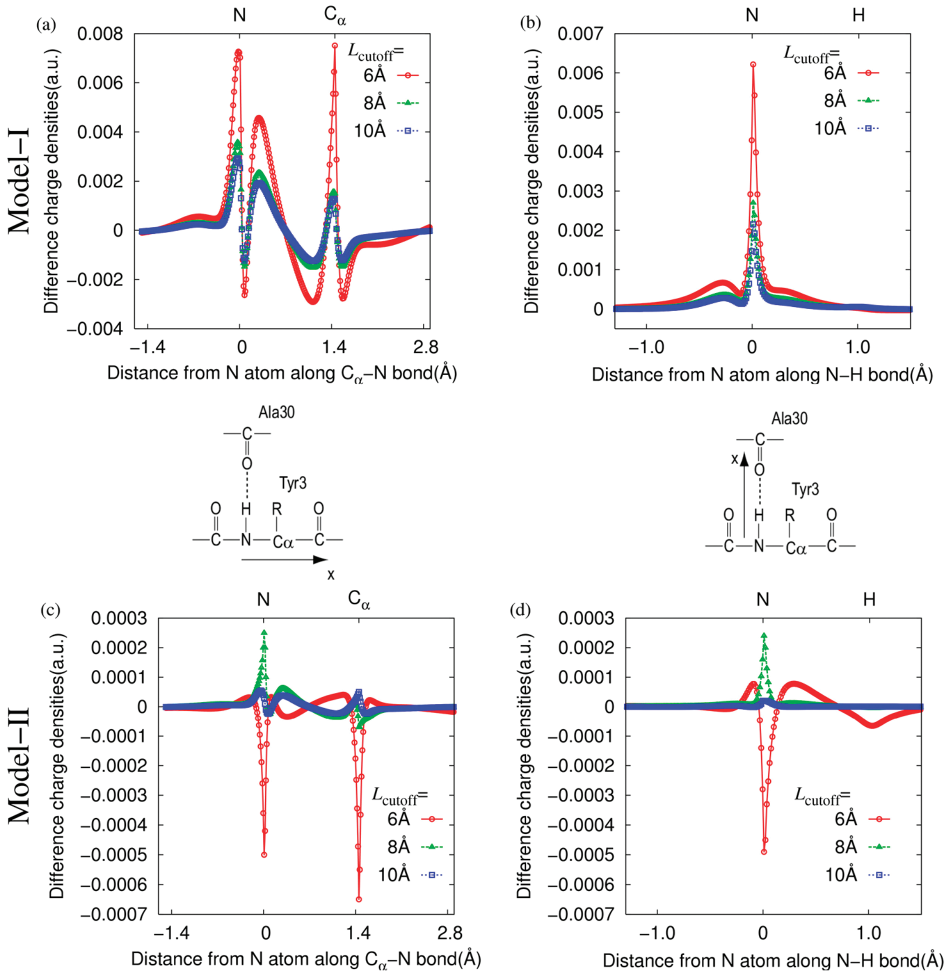


***Figure 8.*** Difference between charge densities determined by FMO1(merged) and conventional *ab initio* calculations with CSGT (6-31G(d)) ($\Delta\rho = \rho_{FMO1(merged)} - \rho_{conventional}$). $\Delta\rho$ of model I (a) along the $C_\alpha$−N bond and (b) along the N−H bond. $\Delta\rho$ of model II (c) along the $C_\alpha$−N bond and (d) along the N−H bond.

density, $\Delta\rho$, at $L_{cutoff}$ = 6, 8, and 10 Å for $C_\alpha$−N and N−H bonds (shown in Figure 8) determined by FMO1(merged) (both models I and II) and conventional *ab initio* calculations are also in excellent agreement. (Note that the scale of the vertical axis in Figure 8 is much smaller than that in Figure 7.) However, it is worth noting that the cutoff-distance dependence of charge distribution is significantly different in the cases of models I and II. The charge distribution calculated with model I strongly depends on $L_{cutoff}$. Increasing $L_{cutoff}$ from 6 to 8 Å substantially improves the descriptions of the charge densities of $C_\alpha$ and N atoms, whereas the improvement is relatively small with increasing $L_{cutoff}$ from 8 to 10 Å. In contrast, the distribution of charge density with model II is almost independent of $L_{cutoff}$ from 6 to 10 Å. This result indicates that a small $L_{cutoff}$ is good enough to reproduce the charge-density distribution in the case of model II. This is because the potential used in model II is calculated from the more realistic density distributions instead of point charges used in model I.

**4.6. Electrostatic Potential.** The charge-density distribution is mainly determined by ESP. The dependence of ESP on the value of $L_{cutoff}$ was therefore assessed. Around Tyr3, the ESP obtained with FMO1(merged) (Figure 9a−f) is compared to that obtained with conventional *ab initio* calculations (Figure 9g). It is clear that ESP is well reproduced in the proximity of Tyr3. However, large differences in ESP determined by FMO1(merged) and the conventional *ab initio* method are found in the right-hand region far from Tyr3 for $L_{cutoff}$ = 6 Å (Figure 9a and b). The ESP is greatly improved around the middle part on the right as the $L_{cutoff}$ increases from 6 to 8 Å (Figure 9c and d). The ESP for $L_{cutoff}$ = 10 Å (Figure 9e and f) is almost indistinguishable from that for conventional *ab initio* calculations (Figure 9g)). Moreover, the differences between model I (Figures 9a, c, and e) and model II (Figures 9b, d, and f) are not obvious.

The differences in ESP ($\Delta V = V_{FMO1(merged)} - V_{conventional}$) along the $C_\alpha$−N and N−H bonds for Tyr3 in the $\beta$-sheet are shown in Figure 10. Similar results to those described above are obtained for the differences between ESP determined by the FMO1(merged) method and the conventional *ab initio* method (Figure S1, Supporting Information). The change in the color in the peripheral area of the polypeptide in Figure 9e and f (with increasing $L_{cutoff}$ from 8 to 10 Å) corresponds to the change in the plateau height in Figure 10. The plateau-height change is due to the neutrality of the system with $L_{cutoff}$ = 10 Å; i.e., the systems with $L_{cutoff}$ = 6 or 8 Å have a positive charge due to the existence of His2 (see Figure 5h), while the system with $L_{cutoff}$ = 10 Å is neutral because of the existence of Glu25. The plateau-height

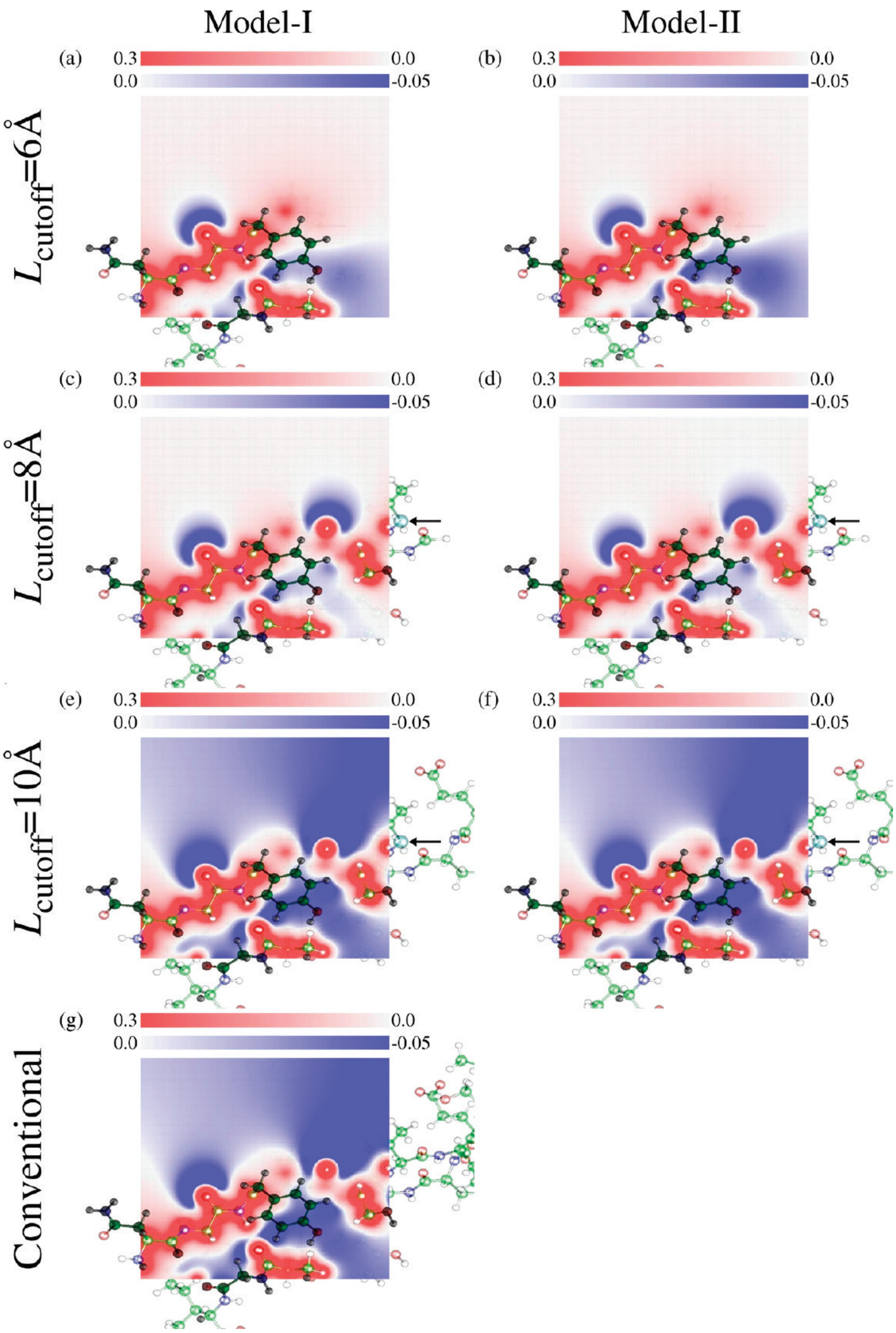


*Figure 9.* Electrostatic potential on the surface specified by three atoms (C, O, and N) of the peptide bond between His2 and Tyr3 in the $\beta$-sheet (15.7 × 15.7 Å$^2$). The magnitude of the ESP is expressed in proportion to the intensity of the colors (red for positive values and blue for negative values), which are indicated as the gradation of the colors at the top of each panel. The unit is $e/(a_0)^3$ where $a_0$ is the Bohr radius. All ESP values are in atomic units. FMO1(merged) results obtained using model I with CSGT (6-31G(d)) with an $L_{\mathrm{cutoff}}$ of (a) 6 Å, (c) 8 Å, and (e) 10 Å and model II with CSGT (6-31G(d)) with an $L_{\mathrm{cutoff}}$ of (b) 6 Å, (d) 8 Å, and (f) 10 Å. (g) Conventional *ab initio* calculation (6-31G(d)) results. One of the border atoms at the fragment boundary is represented as a light green transparent sphere and indicated by an arrow for each panel.

change, however, merely induces a shift in the potential near Tyr3. Consequently, the charge distribution in Figure 8 only changes slightly when $L_{\mathrm{cutoff}}$ increases from 8 Å and 10 Å.

**4.7. Magnetic Susceptibility.** The effects of magnetic susceptibility are considered next. Previous studies[58–62] using *ab initio* calculations and some empirical approaches to determine the contribution of magnetic susceptibilities to chemical shifts have revealed that this contribution mostly arises from the ring current near conjugated groups such as benzene and carbonyl bonds. These effects have a large influence on the shielding tensor when the distance between observed atoms and the functional groups is smaller than the size of the functional groups. Accordingly, the ring-current effect was investigated by calculating the effect of the phenyl group by performing an additional calculation where a phenyl group is replaced by a hydrogen atom. (The position of the hydrogen atom is optimized by AM1.)

As shown in Figure 11, the change in the isotropic shielding constants (more than 0.1 ppm) due to the elimination of the phenyl group of Phe extends about 10 Å from

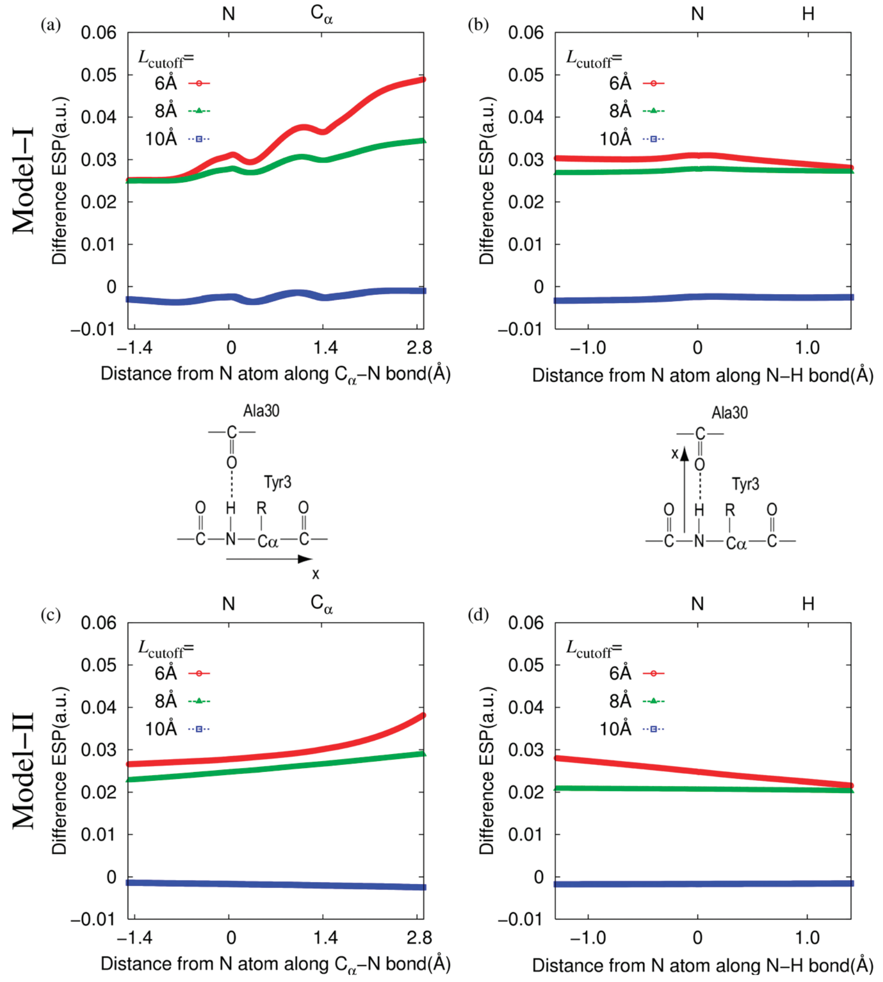


***Figure 10.*** Difference between ESPs determined by FMO1(merged) and conventional *ab initio* calculations with CSGT (6-31G(d)) ($\Delta V = V_{\mathrm{FMO1(merged)}} - V_{\mathrm{conventional}}$). $\Delta V$ of model I (a) along the $C_\alpha$−N bond and (b) along the N−H bond. $\Delta V$ of model II (c) along the $C_\alpha$−N bond and (d) along the N−H bond.

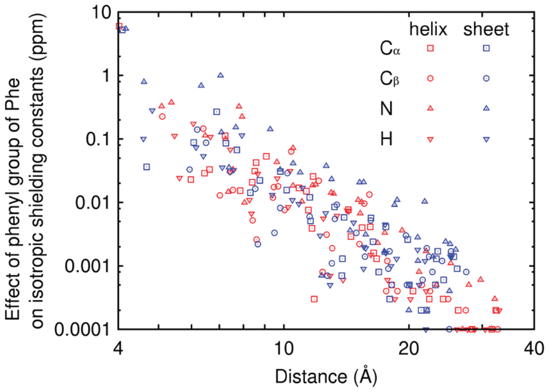


***Figure 11.*** Absolute values of difference in isotropic shielding constants in the cases with or without the phenyl group of Phe in the α-helix and in the β-sheet. Distances from the center of mass of six carbon atoms of the phenyl group to $C_\alpha$, $C_\beta$, N, and H are shown as abscissa.

the center of the phenyl group. To confirm that the effect is not due to the difference in charge density in the cases with and without the phenyl group, the difference between the charge-density distribution with and without the phenyl group was assessed. Contours of the absolute values of the differences in electron densities in the cases with or without the phenyl group of Phe in the α-helix are shown in Figure 12. This figure indicates that the change in the charge density is localized near the phenyl group and extends over a distance of less than 10 Å. A similar result is obtained in the case of the β-sheet (Figure S2, Supporting Information). Thus, the difference in the isotropic shielding constants in Figure 11 at around 10 Å from the phenyl group is not due to the change in the charge density but mainly due to the ring current. Typically, a distance of about 10 Å from the phenyl group to the atoms in the backbone corresponds to an $L_{\mathrm{cutoff}}$ of 8 Å. Thus, an $L_{\mathrm{cutoff}}$ of 8 Å is an adequate distance to achieve our target accuracy of 0.1 ppm.

**4.8. FMO Boundary.** The ability to predict the electronic structure near the boundary region is discussed in the following. An artificially determined boundary may not reproduce the electronic structures accurately, considerably influencing the shielding tensor of the atoms in the vicinity of the boundary. The boundary region was thus examined by analyzing the difference between charge densities determined by the FMO1(merged) and the conventional *ab initio* method. As can be seen in Figure 6, charge densities by model I have much larger deviations in the case of conventional *ab initio* calculations compared to those obtained with model II. For example, model I (Figure 6c)

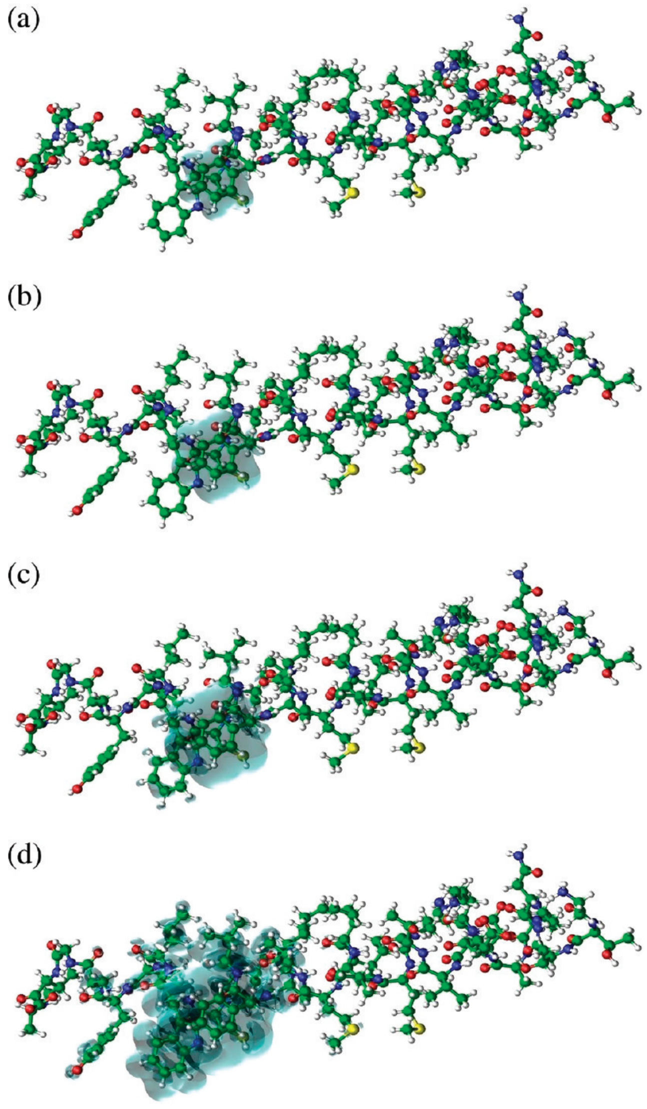


***Figure 12.*** Absolute values of differences in electron densities in the cases with or without the phenyl group of Phe in the α-helix. Contour values are (a) 0.01, (b) 0.001, (c) 0.0001, and (d) 0.00001 $e/(a_0)^3$, where $a_0$ is the Bohr radius.

shows the largest differences in charge densities around the boundary, i.e., about 0.0044 $e/(a_0)^3$, whereas model II (Figure 6d) gives values of about 0.0007 $e/(a_0)^3$. The large errors in charge density near the boundary in the case of model I are due to the point-charge approximation. However, because model II uses the density matrix of surrounding monomers, its quality is superior to that of model I, providing a more accurate shielding tensor.

## 5. Summary

A new *ab initio* method for calculating NMR chemical shifts—named “FMO1(merged)”—has been developed. Chemical shifts for the α-helix and *β*-sheet are calculated by using the point-charge environment (referred to as model I) and the ESP of other fragments without approximation (model II). NMR shifts determined with model I agree well with those calculated by conventional *ab initio* methods (GAIO and CSGT). However, the accuracy of the chemical shifts determined by model I is similar to that of those determined by FMO1(merged) without the $V_{\mu\nu}^{Q(L_{\text{cutoff}})}$ term, where the ESPs of other fragments are completely neglected. Much better results were obtained with model II; that is, it provides an accurate description of the ESP and charge density around all the atoms, which results in a much better accuracy in the calculation of NMR chemical shifts. The conventional *ab initio* NMR results were reproduced using FMO1(merged)/model II with adequate cutoff distances (i.e., $L_{\text{cutoff}} \geq 8$ Å). This adequate value for the cutoff distance was confirmed for $^{13}C_\alpha$, $^{13}C_\beta$, $^{15}N$, and $^{1}H$ atoms in α-helix and *β*-sheet polypeptides with the 6-31G(d) and 6-311G(d,p) basis sets. The proposed method extends the use of the FMO method.

**Acknowledgment.** We thank Prof. M. Fujihira and Dr T. Kohno for discussion and encouragement. All the calculations were performed on the TSUBAME Grid Cluster in Tokyo Institute of Technology.

**Supporting Information Available:** Tables of results calculated with FMO1(merged)/model I and FMO1(merged)/model II by using CSGT at the 6-311G(d,p) level, FMO1(merged)/model I by using GIAO at the 6-31G(d) and 6-311G(d,p) levels, and FMO1(dimer) without the $V_{\mu\nu}^{Q(L_{\text{cutoff}})}$ term by using CSGT at the 6-31G(d) level, a figure for the difference in ESP between FMO1(merged) and conventional *ab initio* calculations by using CSGT, a figure for the calibration of the effect of ring current. This material is available free of charge via the Internet at http://pubs.acs.org.

### References

(1) Ditchfield, R. *Mol. Phys.* **1974**, *27*, 789–807.

(2) Wolinski, K.; Hinton, J. F.; Pulay, P. *J. Am. Chem. Soc.* **1990**, *112*, 8251–8260.

(3) Keith, T. A.; Bader, R. F. W. *Chem. Phys. Lett.* **1993**, *210*, 223–231.

(4) One of the examples of ab initio calculations of a large peptide molecule is found in Wang, B.; Miskolzie, M.; Kotovych, G.; Pulay, P. *J. Biomol. Struct. Dyn.* **2002**, *20*, 71–79.

(5) Warshel, A.; Karplus, M. *J. Am. Chem. Soc.* **1972**, *94*, 5612–5625.

(6) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227–249.

(7) Cui, Q.; Karplus, M. *J. Phys. Chem. B* **2000**, *104*, 3721–3743.

(8) Ishida, T. *Biochemistry* **2006**, *45*, 5413–5420.

(9) He, X.; Wang, B.; Merz, K. M. *J. Phys. Chem. B* **2009**, *113*, 10380–10388.

(10) Sebastiani, D.; Rothlisberger, U. *J. Phys. Chem. B* **2004**, *108*, 2807–2815.

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

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

(13) Field, M. J.; Bash, P. A.; Karplus, M. *J. Comput. Chem.* **1990**, *11*, 700–733.

(14) Gascón, J. A.; Sproviero, E. M.; Batista, V. S. *J. Chem. Theory Comput.* **2005**, *1*, 674–685.

(15) Gascón, J. A.; Sproviero, E. M.; Batista, V. S. *Acc. Chem. Res.* **2006**, *39*, 184–193.

(16) Svensson, M.; Humbel, S.; Froese, R. D. J.; Matsubara, T.; Sieber, S.; Morokuma, K. *J. Phys. Chem.* **1996**, *100*, 19357–19363.

(17) Hall, K. F.; Vreven, T.; Frisch, M. J.; Bearpark, M. J. *J. Mol. Biol.* **2008**, *383*, 106–121.

(18) Hua, W.; Fang, T.; Li, W.; Yu, J.-G.; Li, S. *J. Phys. Chem. A* **2008**, *112*, 10864–10872.

(19) Rahalkar, A. P.; Ganesh, V.; Gadre, S. R. *J. Chem. Phys.* **2008**, *129*, 234101.

(20) He, J.; Di Paola, C.; Kantorovich, L. *J. Chem. Phys.* **2009**, *130*, 144104.

(21) Söderhjelm, P.; Ryde, U. *J. Phys. Chem. A* **2009**, *113*, 617–627.

(22) Xie, W.; Orozco, M.; Truhlar, D. G.; Gao, J. *J. Chem. Theory Comput.* **2009**, *5*, 459–467.

(23) Leverentz, H. R.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 1573–1584.

(24) Suárez, E.; Díaz, N.; Suárez, D. *J. Chem. Theory Comput.* **2009**, *5*, 1667–1679.

(25) Gordon, M. S.; Mullin, J. M.; Pruitt, S. R.; Roskop, L. B.; Slipchenko, L. V.; Boatz, J. A. *J. Phys. Chem. B* **2009**, *113*, 9646–9663.

(26) Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M. *Chem. Phys. Lett.* **1999**, *313*, 701–706.

(27) Fedorov, D. G.; Kitaura, K. *J. Chem. Phys.* **2005**, *122*, 054108.

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

(29) Fedorov, D. G.; Kitaura, K. *J. Chem. Phys.* **2004**, *121*, 2483–2490.

(30) Fedorov, D. G.; Kitaura, K. *J. Phys. Chem. A* **2007**, *111*, 6904–6914.

(31) Fedorov, D. G.; Kitaura, K. *The fragment molecular orbital method: practical applications to large molecular systems*; CRC Press: Boca Raton, FL, 2009.

(32) Nakanishi, I.; Fedorov, D. G.; Kitaura, K. *Proteins: Struct., Funct., Bioinf.* **2007**, *68*, 145–158.

(33) He, X.; Fusti-Molnar, L.; Cui, G.; Merz, K. M. *J. Phys. Chem. B* **2009**, *113*, 5290–5300.

(34) Yoshida, T.; Fujita, T.; Chuman, H. *Curr. Comput.-Aided Drug Des.* **2009**, *5*, 38–55.

(35) Takematsu, K.; Fukuzawa, K.; Omagari, K.; Nakajima, S.; Nakajima, K.; Mochizuki, Y.; Nakano, T.; Watanabe, H.; Tanaka, S. *J. Phys. Chem. B* **2009**, *113*, 4991–4994.

(36) Sawada, T.; Fedorov, D. G.; Kitaura, K. *Int. J. Quantum Chem.* **2009**, *109*, 2033–2045.

(37) Sekino, H.; Matsumura, N.; Sengoku, Y. *Comput. Lett.* **2007**, *3*, 423–430.

(38) Gao, Q.; Yokojima, S.; Kohno, T.; Ishida, T.; Fedorov, D. G.; Kitaura, K.; Fujihira, M.; Nakamura, S. *Chem. Phys. Lett.* **2007**, *445*, 331–339.

(39) Harris, R. K.; Becker, E. D.; Cabral de Menezes, S. M.; Goodfellow, R.; Granger, P. *Pure Appl. Chem.* **2001**, *73*, 1795–1818.

(40) Yokojima, S.; Gao, Q.; Nakamura, S. *AIP Conf. Proc.* **2009**, *1102*, 164–167.

(41) Harris, R. K.; Jackson, P.; Merwin, L. H.; Say, B. J.; Hägele, G. *J. Chem. Soc., Faraday Trans. 1* **1988**, *84*, 3649–3672.

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

(43) Yates, J. R.; Pham, T. N.; Pickard, C. J.; Mauri, F.; Amado, A. M.; Gil, A. M.; Brown, S. P. *J. Am. Chem. Soc.* **2005**, *127*, 10216–10220.

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

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

(46) Ishida, T.; Fedorov, D. G.; Kitaura, K. *J. Phys. Chem. B* **2006**, *110*, 1457–1463.

(47) Sitkoff, D.; Case, D. A. *Prog. NMR Spectr.* **1998**, *32*, 165–190.

(48) Teller, D. C.; Okada, T.; Behnke, C. A.; Palczewski, K.; Stenkamp, R. E. *Biochemistry* **2001**, *40*, 7761–7772.

(49) Jain, R. K.; Ranganathan, R. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 111–116.

(50) Case, D. A.; Darden, T. A.; Cheatham, T. A., III; 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 8*; University of California: San Francisco, 2004.

(51) Wang, J.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049–1074.

(52) Stewart, J. J. P. *Int. J. Quantum Chem.* **1996**, *58*, 133–146.

(53) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.

(54) Stewart, J. J. P.; *MOPAC2007*, v. 1.0; Fujitsu Limited: Tokyo, 2007.

(55) Mulliken, R. S. *J. Chem. Phys.* **1955**, *23*, 1833–1840.

(56) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Puis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347–1363.

(57) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A.; Vreven, T., 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.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng,

C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, Revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.

(58) Ösapay, K.; Case, D. A. *J. Am. Chem. Soc.* **1991**, *113*, 9436–9444.

(59) Case, D. A. *J. Biomol. NMR* **1995**, *6*, 341–346.

(60) Heine, T.; Corminboeuf, C.; Seifert, G. *Chem. Rev.* **2005**, *105*, 3889–3910.

(61) Johnson, C. E.; Bovey, F. A. *J. Chem. Phys.* **1958**, *29*, 1012–1014.

(62) Giessner-Prettre, C.; Pullman, B. *Q. Rev. Biophys.* **1987**, *20*, 113–172.

CT100006N

JCTC Journal of Chemical Theory and Computation

# *Erratum*

**E2 and $S_N2$ Reactions of $X^-$ + $CH_3CH_2X$ (X = F, Cl); an *ab Initio* and DFT Benchmark Study.** [*J. Chem. Theory Comput. 4,* 929–940 (2008)]. By A. Patrícia Bento, Miquel Solà, and F. Matthias Bickelhaupt*.

Page 933. In Table 1, the CCSD(T) value of the **1bPC** species at the CBS limit obtained from two-point fits (aug-cc-pVTZ and aug-cc-pVQZ) is −34.27 kcal/mol and not −37.39 kcal/mol as indicated in the original paper. Conclusions are not affected, because they are based on the aug-cc-pVQZ values and not on the CBS extrapolations.

CT100103M

10.1021/ct100103m
Published on Web 03/12/2010

Published on Web 03/12/2010