

## Hydrogen Bonding Cooperativity in polyQ $\beta$ -Sheets from First Principle Calculations

Giulia Rossetti,<sup>†,‡,§</sup> Alessandra Magistrato,<sup>\*,†</sup>  
Annalisa Pastore,<sup>||</sup> and Paolo Carloni<sup>†,‡,§</sup>

Statistical and Biological Physics Sector, International School for Advanced Studies (SISSA-ISAS) and CNR-IOM-DEMOCRITOS National Simulation Center, Trieste, Italy, Via Beirut 2-4, Trieste, Italy, German Research School for Simulation Science, FZ-Juelich and RWTH, Germany, Italian Institute of Technology—SISSA Unit, Via Beirut 2-4, Trieste, Italy, and National Institute for Medical Research, The Ridgeway London, NW71AA, United Kingdom

Received September 09, 2009

**Abstract:** Polyglutamine  $\beta$ -sheet aggregates are associated with the derangement of Huntington's disease. The effect of cooperativity of the H-bond network formed by both backbone and side chain groups is expected to be important for the structure and energetics of the aggregates. So far, no direct description and/or quantification of the effect is yet available. By performing DFT and hybrid DFT/MM simulations of polyglutamine  $\beta$ -sheet structures *in vacuo* and in aqueous solution, we observe that the cooperativity of glutamine side chains affects both the directions perpendicular and parallel to the backbone. This behavior is not usually observed in  $\beta$  sheets and may provide significant extra-stabilization together with explaining some of the unique properties of polyglutamine aggregation.

Huntington's and other neurodegenerative diseases depend on the abnormal expansion of polyglutamine (polyQ) tracts in proteins which form aggregates rich in  $\beta$  sheets associated with neurodegeneration.<sup>1–6</sup> The glutamine side chain is similar to the backbone unit. Thus, polyQ tracts can form particular  $\beta$  strands stabilized by a hydrogen bond (HB) net involving both the backbone and the side chains.<sup>5–7</sup> The presence of a

cooperative effect (CE) on this peculiar HB net may play a role in the misfolding and aggregation of polyQ.<sup>5</sup> The CE in hydrogen bonding is very important for both the structure and the energetics of polypeptide systems.<sup>8</sup>

The extra-structural stability of polyQ aggregates due to the CE is related to the number of HBs formed between the backbone and side chains.<sup>9</sup> Nevertheless, the conclusions so far were achieved by classical molecular dynamics calculations that cannot answer the critical issue of how to deal with electronic polarizability. This can be described by first principle methods, which have in fact already been applied in the study of CE on polypeptides, including polyQ chains.<sup>10–18</sup> However, the crucial role of Q side chain HBs on the CE has not been investigated so far by first principles approaches.

Here, we perform first principles DFT-PBE<sup>19–21</sup> calculations on polyQ peptides of increasing complexity, assembled in parallel  $\beta$  sheets (Table 1), a structure well characterized from biochemical and theoretical studies [CE turns out to be stronger in parallel  $\beta$  sheets (like the systems considered here) than in antiparallel ones<sup>22</sup>].<sup>14,22–24</sup> Our models ( $N \times n$  hereafter) differed from each other for the number of strands ( $N = 1, 2, 3, 4$ ) and/or for the number of Qs in each strand ( $n = 1, 2, 3, 4$ ) [the models were built using the *HyperChem 8.0* program<sup>25</sup>]. They are terminated by the addition of  $-\text{NCH}_3$  and  $-\text{OCCH}_3$  groups. The resulting 16 models range from 29 to 320 atoms (Table 1, see the footnote for more details on notations). Next, because of the obvious role of solvent and temperature effects on polypeptide conformation,<sup>13</sup> we performed 2 ps of hybrid DFT/MM molecular dynamics calculations on a large system, a  $\beta$ -helix nanotube (8 turns of 20 Q, see Figure S1 in the Supporting Information) in aqueous solution.<sup>29–32</sup> [The structure is characterized by Q residues with  $\varphi$  and  $\psi$  angles of  $-162^\circ$  and  $159^\circ$ .<sup>26</sup> Its coordinates were kindly provided by Dr. A. Lesk. Although  $\alpha$  helices have a low probability of forming *in vivo* with respect to other Q structures,<sup>27,28</sup> they have been investigated here because (1) they have been already investigated by classical MD by us<sup>9</sup> and (2) we provide a qualitative description CE, independently from the peculiarity of these conformations. Quantitative predictions, which would require an investigation on a variety of structures proposed, are beyond the scope of the present investigation.]

Taken together, our calculations suggest that the CE is manifested both by the shortening of HB lengths, increasing the number of HBs involved (**structural** aspect) and by the energy stabilization of H-bonded peptides with respect to the isolated ones (**energetic** aspect): We are going to detail in the following some of the crucial features of our results.

\* Corresponding author e-mail: alema@sissa.it.

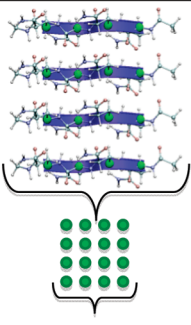
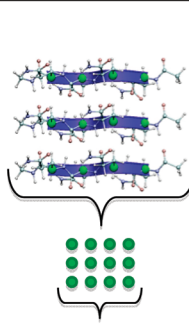
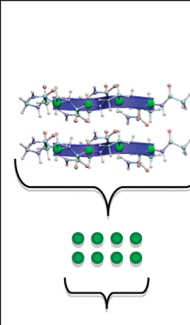
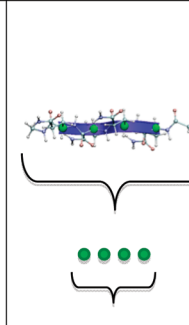
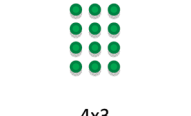




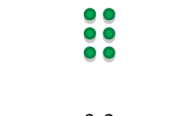
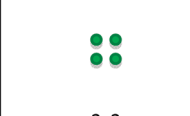





† SISSA and CNR-IOM-DEMOCRITOS.

‡ German Research School for Simulation Science.

§ Italian Institute of Technology—SISSA Unit.

|| National Institute for Medical Research.

**Table 1.** polyQ Peptides of Increasing Complexity, Assembled in Parallel  $\beta$  Sheets<sup>a</sup>

Series $N \times 4$				
Series $N \times 3$				
Series $N \times 2$				
Series $N \times 1$				

<sup>a</sup> Each system studied here is defined in terms of the  $n$  and  $N$  integers, ranging from 1 to 4. The first number counts the Qs in each strand. It defines a group of four systems, each with the same number of Qs per strand, but with a different number of strands (a “series”). The second counts the strands in each system. Thus,  $N \times 4$  indicates systems formed by peptides of 4 Qs ( $1 \times 4$ ,  $2 \times 4$ ,  $3 \times 4$ ,  $4 \times 4$ ),  $N \times 3$  those formed by 3 Qs ( $1 \times 3$ ,  $2 \times 3$ ,  $3 \times 3$ ,  $4 \times 3$ ),  $N \times 2$  those formed by 2 Qs ( $1 \times 2$ ,  $2 \times 2$ ,  $3 \times 2$ ,  $4 \times 2$ ), and  $N \times 1$  those made up of only 1 Q ( $1 \times 1$ ,  $2 \times 1$ ,  $3 \times 1$ ,  $4 \times 1$ ). We built also two other, different  $N \times 3$  series: (A) the  $N \times 3_{SC}$  polyQ series where we varied the side chain conformations and (B)  $N \times 3_{ALA}$ . This is a polyalanine system.

Finally, to prove that such cooperative effects are due only to the peculiarity of polyQ chains, we also considered, as a control study (a) series of models where we varied the initial Q side chain conformations and (b) series of models built with polyalanine.

**Structural Aspects.** CE on a  $\beta$ -sheet system may be present in patterns *perpendicular* to the peptide elongation ( $\perp$ CE) or *parallel* to it (||CE, Figure 1A).<sup>18</sup>

1. The  $\perp$ CE is manifested (a) by a decrease of HB length with an increasing number of piling strands, and (b) by HBs at the center of the pile shorter than in the rim.<sup>18,22</sup>

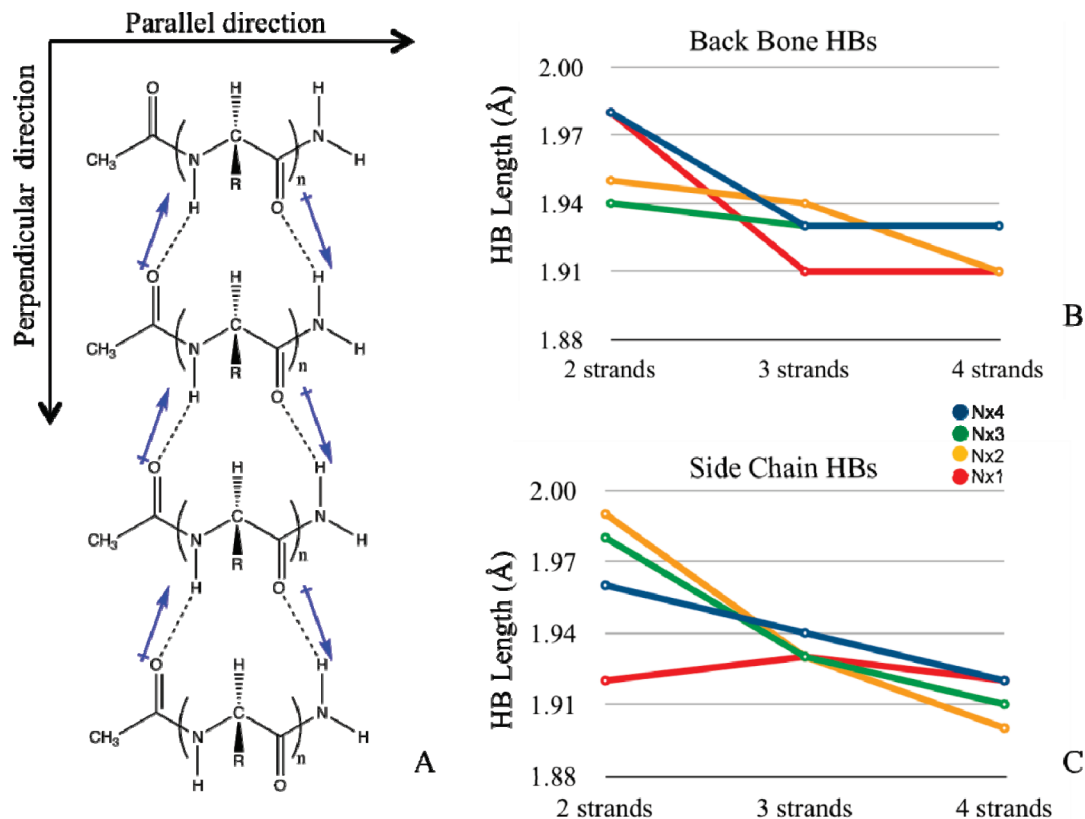
In all the series considered, HB distances decreased with an increasing number of piling strands in both the backbone and the side chains (effect a in Figure 1B,C).

In addition, HB lengths turned out to be shorter at the center of H-bonded chains than at the rim, in the case where at least three HBs are piled up in the perpendicular direction ( $N = 4$ ; effect b in Figures 2–4A and Figure S2, Supporting Information). This feature was observed both for the side chains (Figure 2) and the backbone (Figures 3, 4A, S2).

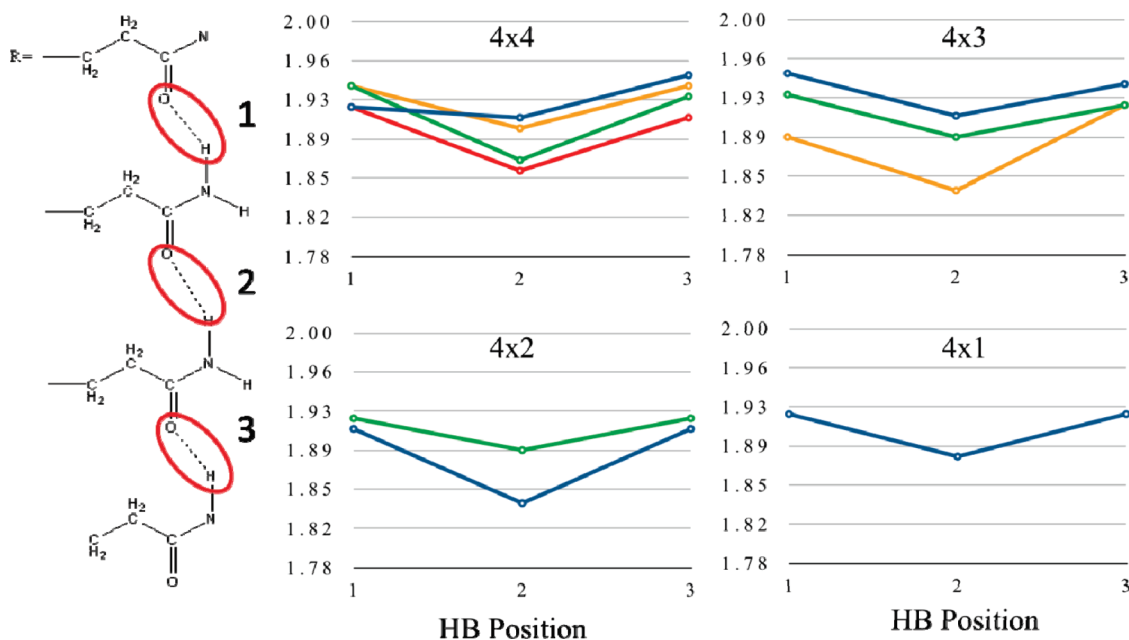
For the backbone, the trend was however observed only when taking averages: the inner HBs turned out not always to be the shortest of the column (Figure 3) [with the term “column”, we indicate the HB chain in the perpendicular direction]. This fact

can be explained, at least in part, by the polarization of the dipoles associated with the HB functionalities ( $C=O \cdots N-H$ ) of both the backbone and side chains. It has been already observed that the backbone dipoles along the same column of  $\beta$  strands have the same orientations (in contrast to those of the adjacent column) and can therefore sum up increasing the polarization of the systems.<sup>22</sup> However, in the case of polyQ  $\beta$ -strands, the glutamine side chains counterbalance this polarization, affecting the inner HBs (Figure 3). Therefore, in the columns where the HB dipole orientations were enhanced by similar side chain HB dipole orientations, a CE is present—the shorter HB was the one in the center of the column; on the other hand, when neighboring side chain columns had HB dipoles oriented in opposite directions (with respect to the column considered), the inner HB was not the shortest of the column (Figure 3).

To prove this conclusion, we perform the same calculations on the  $N \times 3$  polyQ series varying the side chain conformations ( $N \times 3_{SC}$  hereafter). Here, glutamine side chains are twisted in such a way they are not able to establish HBs; thus only backbone HBs are present. As expected, we found both types of  $\perp$ CE (effects a and b). However, remarkably,  $\perp$ CE-type b is not affected by side chain HB dipole orientations due to the absence of side-chain HBs. Thus, backbone HB lengths turned



**Figure 1.** (A) Parallel ( $\parallel$ ) and perpendicular ( $\perp$ ) directions of peptides elongation. (B, C) Structural aspects of CE: (B) Backbone CE ( $\perp$ CE-effect a): mean values of HB lengths of the backbone atoms versus the number of strands for each series of  $n$  Q. (C) Side chain CE ( $\perp$ CE-effect a): mean values of HB lengths for the side chain atoms versus the number of strands for each series of  $n$  Q.



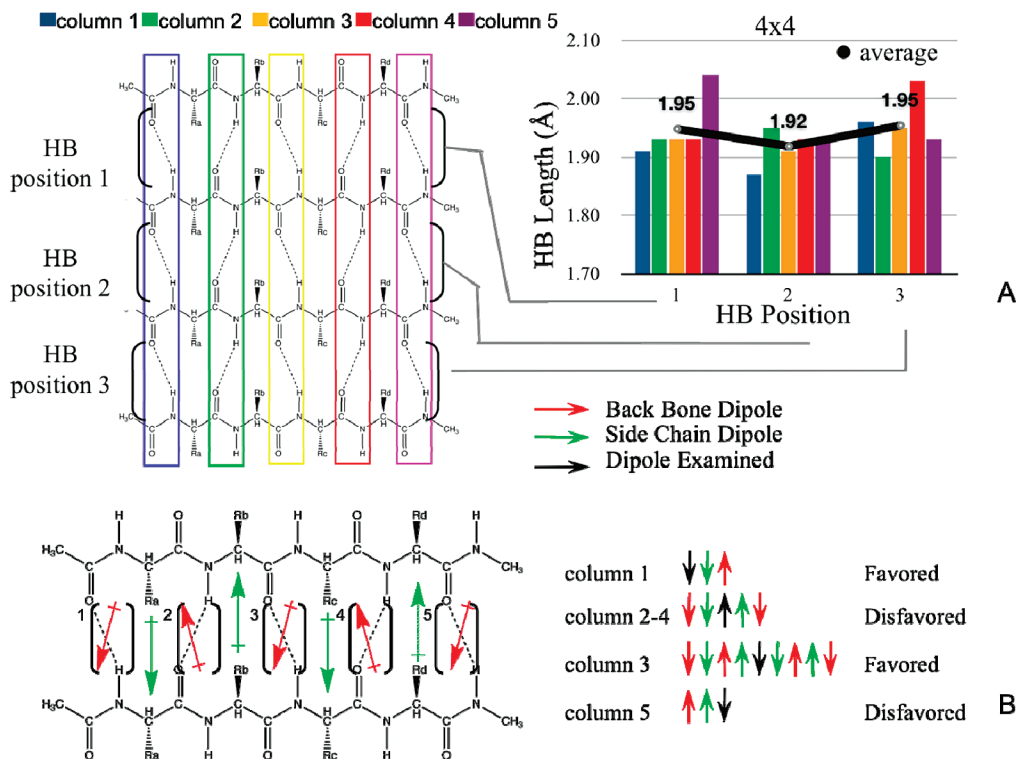
**Figure 2.**  $\perp$ CE-effect b, in the Q side chains: systems  $4 \times 4$ ,  $4 \times 3$ ,  $4 \times 2$ , and  $4 \times 1$ . HB length versus HB position.

out to be shorter at the center of H-bonded chains than at the rim, in each single columns considered, not only taking the average (Figure S3 and Table S2, Supporting Information).

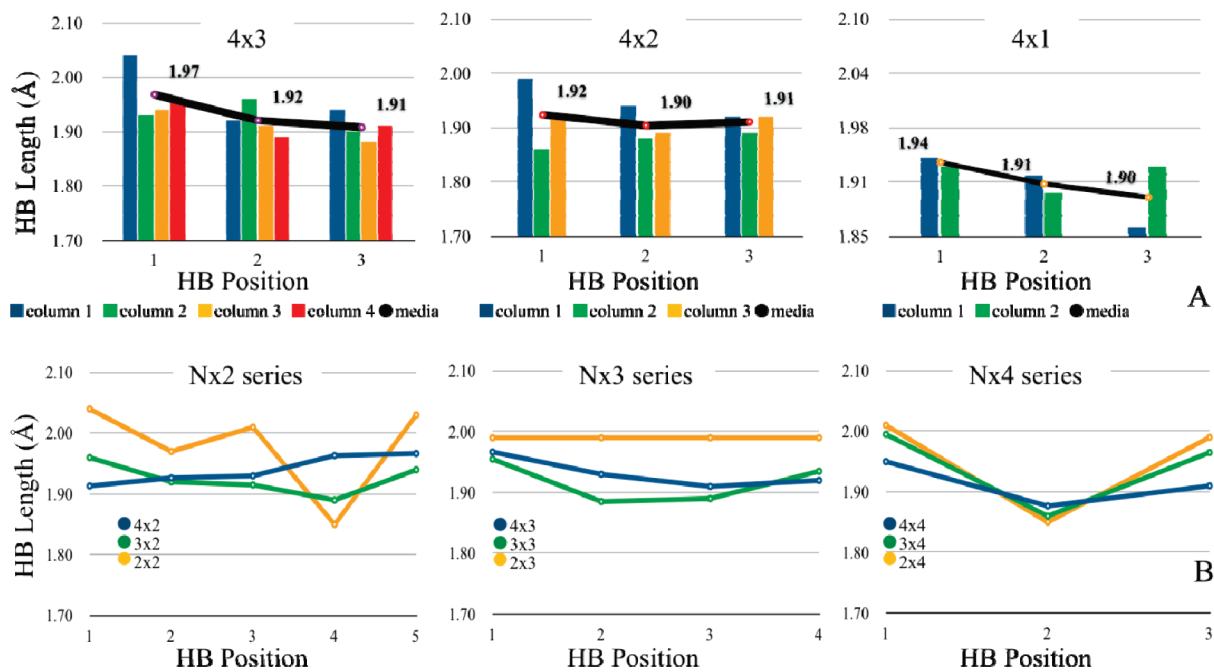
According to this, we have similar results also for  $N \times 3_{\text{ALA}}$ , where there is no possibility for the alanine to form side chain HBs (Figure S3 and Table S2, Supporting Information). Indeed,

we found only a  $\perp$ CE of type b: HB lengths are shorter at the center of H-bonded chains than at the rim, in the case where at least three HBs are piled up in the perpendicular direction ( $N = 4$ ; Figure S3).

2. We observed a  $\parallel$ CE as reflected from shortening of the central HB lengths between two adjacent strands.<sup>18</sup>  $\parallel$ CE is



**Figure 3.**  $\perp$ CE-effect b in system  $4 \times 4$ . (A) In the histograms: HB length of backbone for different positions inside each strand as a function of the position across the different strands. The color of the histogram corresponds to the HBs circled on the top-left picture. The black line represents the mean values over the rows. (B) Orientation of the dipoles associated with the HBs for  $4 \times 4$  ( $4 \times 3$ ,  $4 \times 2$ , and  $4 \times 1$  treated in Supporting Information, Figure S2).



**Figure 4.** (A) Backbone CE in the direction perpendicular to strand elongation ( $\perp$ CE-effect b): systems  $4 \times 3$ ,  $4 \times 2$ , and  $4 \times 1$ . In the histograms: HB length for each column (the position along the strand) versus the HB position (the position perpendicular to the strand direction). The black line represents the mean values over the rows. (B) Backbone  $\perp$ CE: series  $N \times 2$ ,  $N \times 3$ , and  $N \times 4$ . HB length versus HB position.

usually not present in  $\beta$ -sheets because of the alternative orientation of backbone HB dipoles along the strands (Figure 1A).<sup>18,33,34</sup> However, the dipoles associated with the Q side chains add up in a coherent way for the central HBs between two strands (position 2 in  $N \times 2$  series; positions 2 and 3 in  $N$

$\times 3$  series; positions 2, 3, and 4 in  $N \times 4$  series). As a result, the latter turned out to be shorter than those of the rim (Figure 1B). We performed the same calculation on the  $N \times 3_{SC}$  systems, where there is not a contribution of side chains' HB. As expected, no  $\perp$ CE is found (Figure S4a, Supporting Informa-



tion). These results point to the relevance of glutamine side chains for the structure of polyQ systems.

To confirm that such cooperative effects are specific only for polyQ and not a general feature of polypeptide chains, we performed a control study also on a series ( $N \times 3_{\text{ALA}}$ ) of polyalanine systems (Table S2, Supporting Information). As expected, no  $\parallel\text{CE}$  or  $\perp\text{CE}$  type a was found (Figure S4 b, Supporting Information).

Similar conclusions can be drawn by our hybrid QM/MM calculations of the circular  $\beta$ -helix of the polyQ chain, in which the QM region corresponds to  $4 \times 4$ ,  $4 \times 3$ , and  $3 \times 4$ , and the rest of the polyQ tracts as well as the water molecules were included in the MM region ( $\sim 45\,000$  atoms). These systems were labeled as  $4 \times 4_{\text{MIX}}$ ,  $4 \times 3_{\text{MIX}}$ , and  $3 \times 4_{\text{MIX}}$ . Although the trend of HB lengths in the first two systems qualitatively resembled that of the corresponding *in vacuo* models, we have to remark that the HB lengths were larger (Table S3 and Figure S5, Supporting Information). Moreover, the side chains formed mostly HB with the solvent. These differences are probably due to the presence of the solvent and to temperature effects, which are completely neglected in the *in vacuo* calculations. [We further notice that, because of the very short time-scale of our QM/MM simulation, our structural parameters may also not have reached equilibration.] No CE was observed in the last system ( $3 \times 4_{\text{MIX}}$ ), possibly because of the small number of strands.

**Energetic Aspects.** The stabilization energy associated with the formation of HBs between the different strands of the systems *in vacuo* is calculated as follows. [Unfortunately, the stabilization associated with the addition of a Q unit starts with the fourth amino acid unit,<sup>11</sup> so it cannot be investigated here. In fact, the number of amino acids in our systems is never greater than four. This issue must be addressed in a further study.] First, we define the stabilization energy *per strand* ( $\Delta E_N$ ) as the energy associated with the addition of the  $N$ th Q strand to the  $Q_{N-1}$  ( $E_{N \times n}$ ), minus the formation energy of the  $N$  isolated strand.

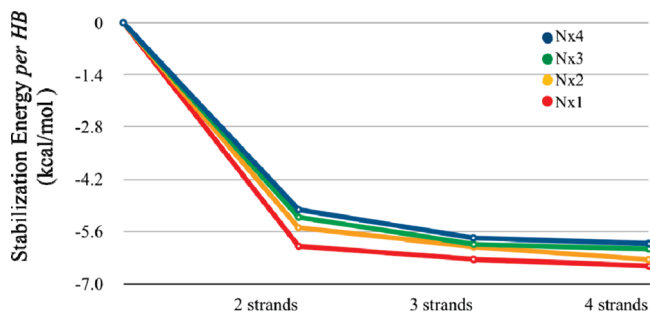
$$\Delta E_N = E_{N \times n} - N \cdot E_{1 \times n}$$

$E_{N \times n}$  is the energy of a system belonging to the  $n$  series and containing  $N$  strands;  $E_{1 \times n}$  is the energy associated with an isolated polyQ strand with  $n$  glutamines. This is the formation energy of a strand constituted by  $n$  glutamines free from long-range effects (i.e., isolated non-interactive strand).

The stabilization energy per hydrogen bond ( $\Delta E_{\text{HB}}$ ) was then defined by dividing  $\Delta E_N$  by the number of hydrogen bonds ( $n_{\text{HB}}$ ) in each system.

$$\Delta E_{\text{HB}} = \Delta E_N / n_{\text{HB}}$$

$\Delta E_{\text{HB}}$  decreased nonlinearly with the number of strands (Figure 5): the variation of  $\Delta E_{\text{HB}}$  in each series is  $\sim 0.8$  kcal/mol, passing from two-strand systems to four-strand systems. This quantity is smaller than the typical DFT-PBE error.<sup>35</sup> However, here, we consider differences of energies in similar systems; thus we can reasonably assume that fortuitous error cancellation errors may increase the accuracy of our calculations.  $\Delta E_{\text{HB}}$  ranged from  $-5.0$  kcal/mol in the smallest system to  $-6.5$  kcal/mol in the larger system ( $4 \times 4$ ), suggesting that a CE



**Figure 5.** Stabilization energy per hydrogen bond ( $\Delta E_{\text{H}}$ ) for the addition of an  $N$ th Q strand to the  $Q_{N-1}$ . The gradual change of  $\Delta E_{\text{H}}$  versus the number of strands showed that the strength of the HBs between layers increases nonlinearly with the number of strands.

effect exists and that for the present systems this is a maximum of 1.5 kcal/mol per HB.

As expected, the stabilization energy depending on CE is smaller for polyA systems with respect to the polyQ, clearly for the absence of side chain HB stabilization. According to this hypothesis, if we compute the CE for the  $N \times 3_{\text{SC}}$  series, where the glutamine side chains are not able to form HBs, we find results comparable with the polyA ones (Table S5, Supporting Information).

In summary, we found that (1) both parallel and perpendicular CEs affect the geometry of polyQ  $\beta$  strands because of the key role of the Q side chains; (2) the formation of cooperative hydrogen bonds stabilized multiple polyQ  $\beta$ -sheet strands with respect to a single isolated strand; (3) within the limitations of the calculations on a single  $\beta$ -stranded structure in a water solution, we suggest that environmental effects on hydrogen bonding CE affects only the magnitude of CE, while the qualitative trend is the same as that found in the *in vacuo* calculation.

**Acknowledgment.** A.P. acknowledges funding from MRC (grant No U117584256).

**Supporting Information Available:** (1) Methods: DFT and DFT/MM calculations. (2) Figure S1, circular  $\beta$ -helix. (3) Figure S2, HB dipole orientations in (a)  $4 \times 3$ , (b)  $4 \times 2$ , and (c)  $4 \times 1$ . (4) Figure S3, (a) backbone CE ( $\perp\text{CE}$ -effect a) in system  $4 \times 3_{\text{SC}}$  and system  $4 \times 3_{\text{ALA}}$ , mean values of HB lengths of the backbone atoms versus the number of strands for each series of  $n$  Q; (b) backbone CE ( $\perp\text{CE}$ -effect b) in the direction perpendicular to strand elongation: system  $4 \times 3_{\text{SC}}$ , system  $4 \times 3_{\text{ALA}}$ . In the histograms: HB length for each column (the position along the strand) versus the HB position (the position perpendicular to the strand direction); the black line represents the mean values over the rows. (5) Figure S4, backbone  $\parallel\text{CE}$ : Series  $N \times 3_{\text{SC}}$ ,  $N \times 3_{\text{ALA}}$ . HB length versus HB position. (6) Figure S5, backbone CE in the direction perpendicular to strand elongation: (a) system  $4 \times 4_{\text{MIX}}$ , (b) system  $4 \times 3_{\text{MIX}}$ . In the histograms: HB length for each column (the position along the strand) versus the HB position (the position perpendicular to the strand direction); the black line represent the mean values over the rows. (7) Table S1, lengths of HBs in backbone and side chains for all the systems studied

with DFT *in vacuo*. (8) Table S2, lengths of HBs in the backbone: series  $N \times 3_{SC}$ ,  $N \times 3_{ALA}$ . (9) Table S3, lengths of HBs in the backbone obtained with DFT/MM MD. (10) Table S4, energies obtained from the DFT calculation *in vacuo*. (11) Table S5, energies obtained from the DFT calculations *in vacuo* for series  $N \times 3_{SC}$ ,  $N \times 3_{ALA}$ . This material is available free of charge via the Internet at <http://pubs.acs.org>.

### References

- (1) Chen, S. M.; Ferrone, F. A.; Wetzel, R. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (18), 11884–11889.
- (2) Davies, S. W.; Turmaine, M.; Cozens, B. A.; DiFiglia, M.; Sharp, A. H.; Ross, C. A.; Scherzinger, E.; Wanker, E. E.; Mangiarini, L.; Bates, G. P. *Cell* **1997**, *90* (3), 537–548.
- (3) Masino, L.; Pastore, A. *Brain Res. Rev.* **2001**, *56* (3–4), 183–189.
- (4) Scherzinger, E.; Sittler, A.; Schweiger, K.; Heiser, V.; Lurz, R.; Hasenbank, R.; Bates, G. P.; Lehrach, H.; Wanker, E. E. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96* (8), 4604–4609.
- (5) Perutz, M. F.; Johnson, T.; Suzuki, M.; Finch, J. T. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91* (12), 5355–5358.
- (6) Perutz, M. F.; Windle, A. H. *Nature* **2001**, *412* (6843), 143–4.
- (7) Klein, F.; Pastore, A.; Masino, L.; Zederlutz, G.; Nierengarten, H.; Ouladabdelghani, M.; Altschuh, D.; Mandel, J.; Trotter, Y. *J. Mol. Biol.* **2007**, *371* (1), 235–244.
- (8) Ludwig, R. *J. Mol. Liq.* **2000**, *84* (1), 65–75.
- (9) Rossetti, G.; Magistrato, A.; Pastore, A.; Persichetti, F.; Carloni, P. *J. Phys. Chem. B* **2008**, *112* (51), 16843–50.
- (10) Horvath, V.; Varga, Z.; Kovacs, A. *THEOCHEM* **2005**, *755* (1–3), 247–251.
- (11) Horvath, V.; Varga, Z.; Kovacs, A. *J. Phys. Chem. A* **2004**, *108*, 6869–6873.
- (12) Improta, R.; Barone, V.; Kudin, K. N.; Scuseria, G. E. *J. Chem. Phys.* **2001**, *114* (6), 2541–2549.
- (13) Scheiner, S.; Kar, T. *J. Phys. Chem. B* **2005**, *109* (8), 3681–3689.
- (14) Tsemekhman, K.; Goldschmidt, L.; Eisenberg, D.; Baker, D. *Protein Sci.* **2007**, *16* (4), 761–4.
- (15) Varga, Z.; Kovacs, A. *Int. J. Quantum Chem.* **2005**, *105* (4), 302–312.
- (16) Viswanathan, R.; Asensio, A.; Dannenberg, J. J. *J. Phys. Chem. A* **2004**, *108* (42), 9205–9212.
- (17) Wiczorek, R.; Dannenberg, J. J. *J. Am. Chem. Soc.* **2003**, *125* (27), 8124–9.
- (18) Zhao, Y. L.; Wu, Y. D. *J. Am. Chem. Soc.* **2002**, *124* (8), 1570–1.
- (19) Benedek, N. A.; Snook, I. K.; Latham, K.; Yarovsky, I. *J. Chem. Phys.* **2005**, *122* (14), 144102.
- (20) Morozov, A. V.; Kortemme, T.; Tsemekhman, K.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (18), 6946–51.
- (21) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77* (18), 3865–3868.
- (22) Koch, O.; Boccola, M.; Klebe, G. *Proteins* **2005**, *61* (2), 310–317.
- (23) Beke, T.; Csizmadia, I.; Perczel, A. *J. Am. Chem. Soc.* **2006**, *128* (15), 5158–5167.
- (24) Perczel, A.; Gaspari, Z.; Csizmadia, I. G. *J. Comput. Chem.* **2005**, *26* (11), 1155–1168.
- (25) *HyperChem 8.0*; Hypercube, Inc.: Gainesville, FL.
- (26) Perutz, M. F.; Finch, J. T.; Berriman, J.; Lesk, A. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (8), 5591–5.
- (27) Sikorski, P.; Atkins, E. *Biomacromolecules* **2005**, *6* (1), 425–32.
- (28) Zanuy, D.; Gunasekaran, K.; Lesk, A. M.; Nussinov, R. *J. Mol. Biol.* **2006**, *358* (1), 330–345.
- (29) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. *Comput. Phys. Commun.* **1995**, *91* (1–3), 43–56.
- (30) *CPMD 3.11.1*; IBM Corp.: Armonk, New York, 1990–2008.
- (31) Laio, A.; VandeVondele, J.; Rothlisberger, U. *J. Chem. Phys.* **2002**, *116* (16), 6941–6947.
- (32) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. *J. Comput. Chem.* **2005**, *26* (16), 1701–1718.
- (33) Hol, W. G.; Halie, L. M.; Sander, C. *Nature* **1981**, *294* (5841), 532–6.
- (34) Kortemme, T.; Ramírez-Alvarado, M.; Serrano, L. *Science* **1998**, *281* (5374), 253–256.
- (35) Piana, S.; Sebastiani, D.; Carloni, P.; Parrinello, M. *J. Am. Chem. Soc.* **2001**, *123* (36), 8730–7.

CT900476E

## Theoretical Investigation of Solvent Effects on Glycosylation Reactions: Stereoselectivity Controlled by Preferential Conformations of the Intermediate Oxacarbenium-Counterion Complex

Hiroko Satoh,<sup>\*,†,‡</sup> Halvor S. Hansen,<sup>†</sup> Shino Manabe,<sup>§</sup> Wilfred F. van Gunsteren,<sup>†</sup> and Philippe H. Hünenberger<sup>\*,†</sup>

Laboratory of Physical Chemistry, Swiss Federal Institute of Technology (ETH), CH-8093 Zürich, Switzerland, National Institute of Informatics, Tokyo 101-8430, Japan, and RIKEN Advanced Science Institute, Saitama 351-0198, Japan

Received March 12, 2010

**Abstract:** The mechanism of solvent effects on the stereoselectivity of glycosylation reactions is investigated using quantum-mechanical (QM) calculations and molecular dynamics (MD) simulations, considering a methyl-protected glucopyranoside triflate as a glycosyl donor equivalent and the solvents acetonitrile, ether, dioxane, or toluene, as well as gas-phase conditions (vacuum). The QM calculations on oxacarbenium-solvent complexes do not provide support to the usual *solvent-coordination hypothesis*, suggesting that an experimentally observed  $\beta$ -selectivity ( $\alpha$ -selectivity) is caused by the preferential coordination of a solvent molecule to the reactive cation on the  $\alpha$ -side ( $\beta$ -side) of the anomeric carbon. Instead, explicit-solvent MD simulations of the oxacarbenium-counterion (triflate ion) complex (along with corresponding QM calculations) are compatible with an alternative mechanism, termed here the *conformer and counterion distribution hypothesis*. This new hypothesis suggests that the stereoselectivity is dictated by two interrelated conformational properties of the reactive complex, namely, (1) the conformational preferences of the oxacarbenium pyranose ring, modulating the steric crowding and exposure of the anomeric carbon toward the  $\alpha$  or  $\beta$  face, and (2) the preferential coordination of the counterion to the oxacarbenium cation on one side of the anomeric carbon, hindering a nucleophilic attack from this side. For example, in acetonitrile, the calculations suggest a dominant B<sub>2,5</sub> ring conformation of the cation with preferential coordination of the counterion on the  $\alpha$  side, both factors leading to the experimentally observed  $\beta$  selectivity. Conversely, in dioxane, they suggest a dominant <sup>4</sup>H<sub>3</sub> ring conformation with preferential counterion coordination on the  $\beta$  side, both factors leading to the experimentally observed  $\alpha$  selectivity.

### 1. Introduction

In recent years, the investigation of the nature, structure, and function of carbohydrates present in biological systems has received increased interest, especially in the context of

glycoscience and chemical biology.<sup>1,2</sup> A major obstacle in the characterization of biologically relevant carbohydrates is the limited availability of pure and structurally well-defined sugar materials; i.e., sugars are usually found in low concentrations and/or in microheterogeneous forms. So far, synthetic chemistry still represents the main access route to oligosaccharides and glycoconjugates with rigorously defined chemical structures.

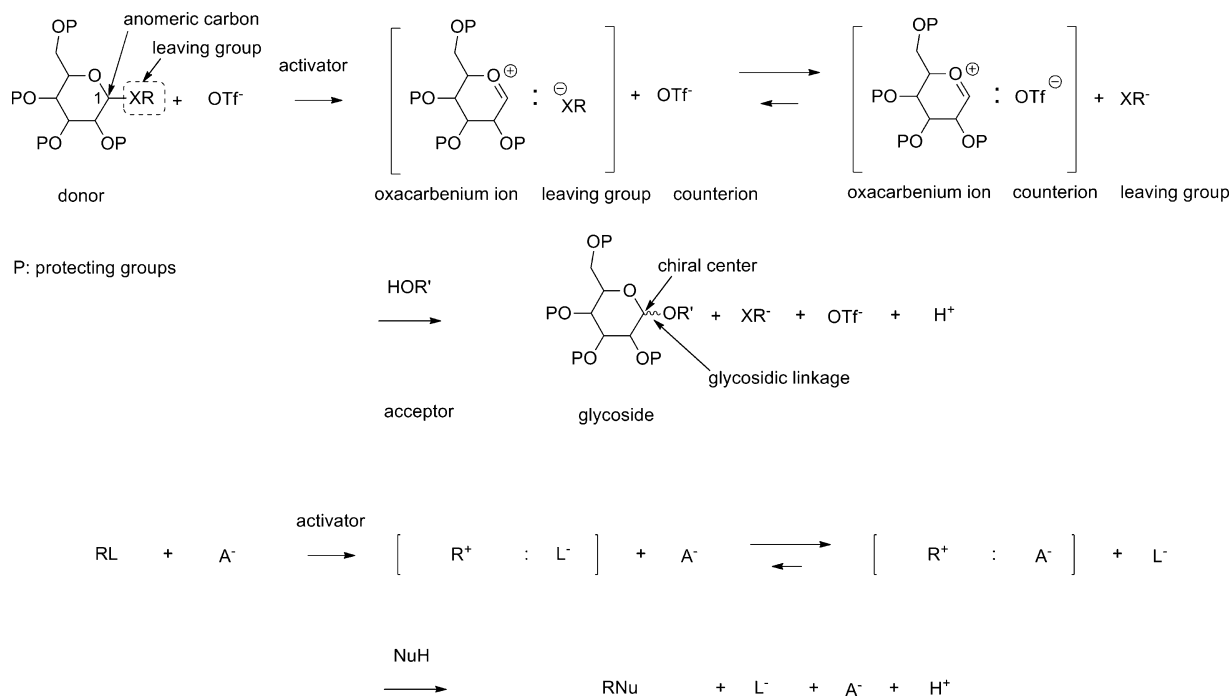
One of the cornerstones of carbohydrate synthesis is the glycosylation reaction, which involves a glycosyl donor

\* Corresponding authors. E-mail: hsatoh@nii.ac.jp (H.S.); phil@igc.phys.chem.ethz.ch (P.H.H.).

<sup>†</sup> Swiss Federal Institute of Technology (ETH).

<sup>‡</sup> National Institute of Informatics.

<sup>§</sup> RIKEN Advanced Science Institute.



**Figure 1.** Generic reaction mechanism of a glycosylation reaction, in a solution containing a specific anion. Top: glycosylation involving a protected pyranoside donor, an alcohol acceptor, and triflate anions in solution. Bottom: generalized version, involving an arbitrary glycosyl donor (RL), an arbitrary nucleophile (NuH), and an arbitrary counterion type ( $A^-$ ).

(electrophile) to be coupled with a glycosyl acceptor (nucleophile) and is promoted by a suitable activator (Figure 1). The exocyclic hydroxyl groups of the donor and of the acceptor that are not involved in the coupling are typically either functionalized or rendered nonreactive by means of protecting groups, while the anomeric carbon (C1) of the donor is functionalized by a leaving group to be substituted by the acceptor. The role of the activator is to assist the departure of this leaving group. Typical leaving groups are imidates, sulfur compounds (thiolates, sulfonates, sulfates), and halogenates. The most common activators are salts of trifluoromethanesulfonate (triflate;  $OTf^-$ ) and a combination of trifluoromethanesulfonate salts with a chalcogen halide. The reaction creates a glycosidic linkage and induces a new chirality at the anomeric carbon. The products usually consist of a mixture of the two possible stereoisomers, i.e., the  $\alpha$ - or  $\beta$ -linkage anomers as defined by IUPAC conventions.<sup>3</sup> Although stereoselective synthetic technologies for oligosaccharides and glycoconjugates have made considerable progress in recent years, including the development of polymer-supported and solid-phase synthesis methodologies,<sup>4–12</sup> a high stereoselectivity is still often difficult to achieve. Since sugar materials that are contaminated with undesirable or indeterminable stereoisomers are less suitable for biological studies, further development of synthetic approaches to control the stereoselectivity of glycosylation reactions is an area of very active research.

Many factors influence the stereoselectivity of a glycosylation reaction, including the choices of the glycosyl donor, leaving group, protecting groups, acceptor, activation system, and solvent, as well as the temperature. Great efforts have been made to gain an understanding of the mechanism of glycosylation reactions and of the relationship between these

factors and the resulting stereoselectivity, in particular *via* synthetic experiments<sup>13–39</sup> and theoretical methods.<sup>40–56</sup>

The reaction mechanism for typical pyranosides is generally assumed to be of the  $SN_1$  type (Figure 1), with a ratio of products under kinetic (rather than thermodynamic) control and transition barriers of a predominantly enthalpic (rather than entropic) nature. This mechanism involves as a first step the (activator assisted) departure of the leaving group ( $L^-$ ) and the formation of an oxocarbenium cation intermediate ( $R^+$ ).<sup>19</sup> This cation benefits from an enhanced stability compared to, e.g., an aliphatic carbocation, generally attributed to the delocalization of the positive charge at the anomeric carbon onto the neighboring ring oxygen atom. For the low to medium polarity organic solvents typically used in glycosylation reactions, a counterion is likely to stay more or less tightly coordinated to this cation. This counterion may be the anionic leaving group  $L^-$  or another type of anion  $A^-$  present in the reaction medium. In this case, the reactive species will be an oxocarbenium–counterion complex intermediate  $[R^+; L^-]$  or  $[R^+; A^-]$ . For example, in the common situation where triflate anions are present, the predominant reactive species will be an oxocarbenium–triflate complex intermediate  $[R^+; OTf^-]$ , which is known as a high reactive glycosylation donor equivalent.<sup>15,17,20,21,23,24,28,32</sup> In a second step, a nucleophile (NuH), typically an alcohol molecule, attacks the anomeric carbon of the intermediate species to form a glycosidic linkage. The nucleophile may attack the oxocarbenium cation from either the  $\alpha$  or the  $\beta$  side, resulting in the formation of either of the two corresponding anomers.

The nature of the solvent is known to represent a key factor in the stereoselectivity of glycosylation reactions.<sup>57–67</sup> For glucopyranosides, for example, the 1,2-*cis*-glucoside ( $\alpha$

**Table 1.** Experimental Results Concerning the Stereoselectivity of Glycosylation Reactions (Figure 1) in Different Solvents (or Solvent Mixtures)<sup>a</sup>

Entry	Donor	Acceptor	Activator	Solvent	$\alpha:\beta$
1		CH <sub>3</sub> OSi(CH <sub>3</sub> ) <sub>3</sub>	TMSOTf	CH <sub>3</sub> CN	16:84
2		CH <sub>3</sub> OSi(CH <sub>3</sub> ) <sub>3</sub>	TMSOTf	Et <sub>2</sub> O	90:10
3		CH <sub>3</sub> OSi(CH <sub>3</sub> ) <sub>3</sub>	TMSOTf	CH <sub>3</sub> CN	22:78
4		CH <sub>3</sub> OSi(CH <sub>3</sub> ) <sub>3</sub>	TMSOTf	Et <sub>2</sub> O	84:16
5			DMTST	CH <sub>2</sub> Cl <sub>2</sub> -CH <sub>3</sub> CN (1:1) % in volume	50:50
6			DMTST	CH <sub>2</sub> Cl <sub>2</sub> -Et <sub>2</sub> O (1:1) % in volume	80:20
7			DMTST	toluene	79:21
8			DMTST	CH <sub>2</sub> Cl <sub>2</sub>	79:21
9				CH <sub>3</sub> CN	18:82
10				Et <sub>2</sub> O	50:50
11				toluene-dioxane (1:1) % in volume	53:47
12				toluene	24:76
13				dioxane-CH <sub>3</sub> CN (1:1) % in volume	24:76
14				CH <sub>2</sub> Cl <sub>2</sub>	44:56

Bn = benzyl      TMSOTf = trimethylsilyl triflate      Tol = *p*-tolyl      MP = *p*-methoxy phenyl  
Phth = phthaloyl      PEG = poly(ethylene glycol)methyl ether      DMTST = dimethylthiomethylsulfonium triflate

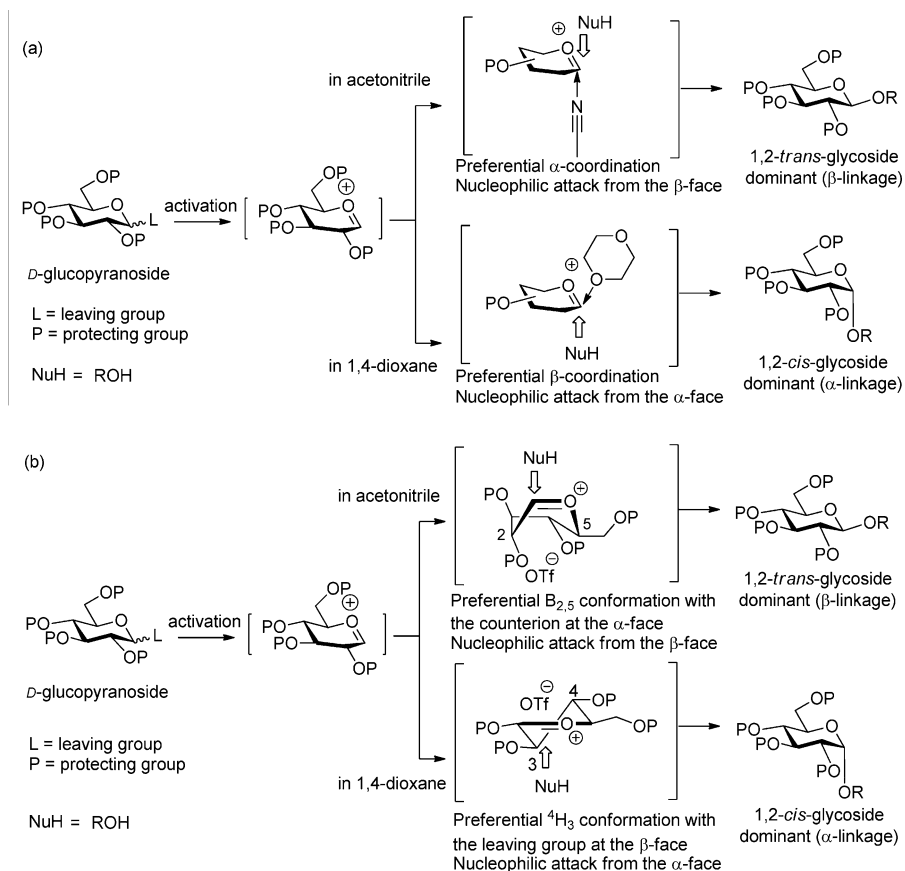
<sup>a</sup> Entries 1–8 are reported from the literature.<sup>61,65</sup> Entries 9–14 correspond to experiments carried out specifically for the present study (see Supporting Information for experimental details).

linkage) is predominantly formed in diethyl ether or in 1,4-dioxane, whereas in acetonitrile, the 1,2-*trans*-glucoside ( $\beta$  linkage) is the major product. The same trend is typically also observed for other pyranosides, namely, that the  $\alpha$ -anomer is predominantly formed in ether or dioxane, whereas the  $\beta$ -anomer is predominantly produced in acetonitrile. The solvent effect dominates the stereoselectivity as long as there is no participating effect of neighboring groups (e.g., participation of a 2-acyl protecting group in the donor, predominantly leading to a 1,2-*trans*-glycosidic linkage irrespective of the solvent<sup>35</sup>).

Some examples from the literature<sup>61,64</sup> demonstrating these solvent effects, along with the results of experiments carried out specifically for the present study, are summarized in

Table 1 (see also the Supporting Information). Entries 1–4<sup>61</sup> show that the stereoselectivity is essentially insensitive to the anomeric configuration of the glycosyl donor, i.e., to the orientation of the leaving group prior to the reaction, providing support for the general assumption of a S<sub>N</sub>1 type glycosylation mechanism. These reactions also evidence a clear  $\beta$ -selectivity in acetonitrile and  $\alpha$ -selectivity in ether. Entries 5–8<sup>64</sup> correspond to reactions on a poly(ethylene glycol)methyl ether (PEG) polymer support. Here, the  $\alpha$ -stereoselectivity observed for an ether/dichloromethane mixture does not differ significantly from that found in solvents usually having little influence on the stereoselectivity (toluene and dichloromethane). However, the acetonitrile/dichloromethane mixture presents a higher proportion of the





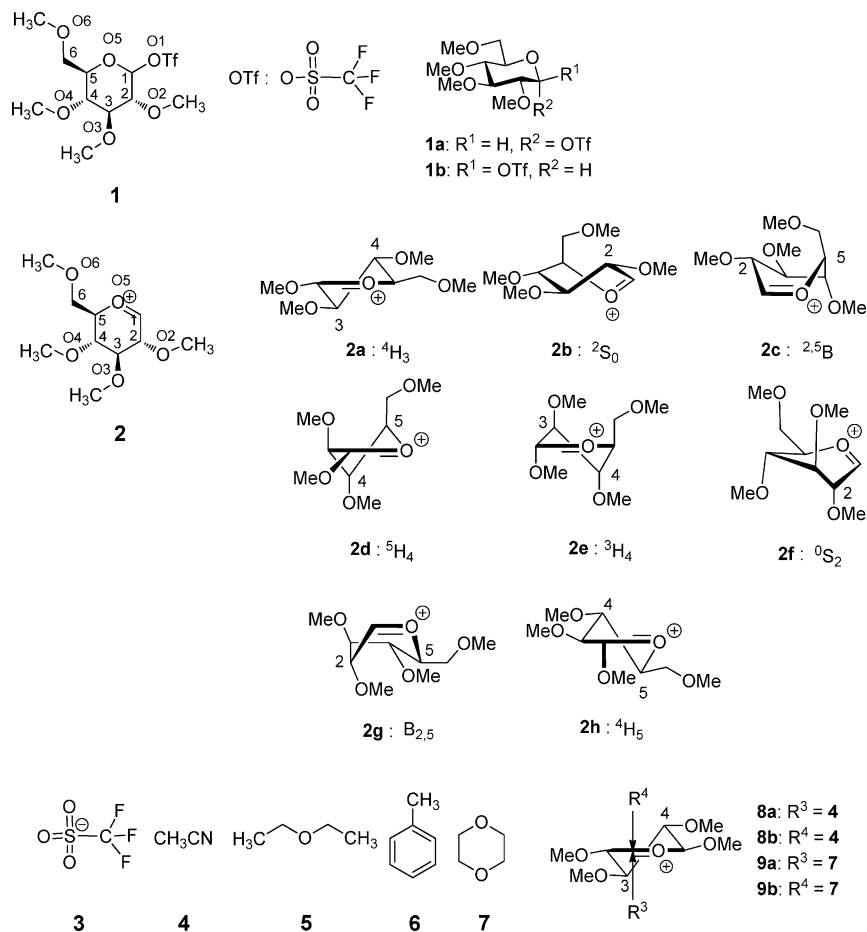
**Figure 2.** Two alternative hypotheses concerning solvent effects in glycosylation reactions: (a) Commonly formulated hypothesis, referred to here as the *solvent coordination hypothesis*; (b) alternative hypothesis formulated on the basis of the present study, referred to here as the *conformer and counterion distribution hypothesis*. A glucopyranoside donor in the solvents acetonitrile and 1,4-dioxane and in the presence of a triflate counterion are selected here to illustrate how the two hypotheses account for the experimentally observed stereoselectivity.

$\beta$ -product, albeit no net  $\beta$ -stereoselectivity. Finally, entries 9–14 (experiments carried out for the present study) evidence a clear  $\beta$ -stereoselectivity in acetonitrile and in an acetonitrile/dioxane mixture (as well as in toluene). The proportion of the  $\alpha$  product is higher in ether and in a dioxane/toluene mixture (as well as in dichloromethane), although no net  $\alpha$ -stereoselectivity is observed. Bearing in mind that the stereoselectivity of glycosylation reactions is affected by many other factors (besides solvent effects), the results presented in Table 1, along with those of other studies,<sup>57–67</sup> suggest a general trend toward  $\beta$ -selectivity in acetonitrile, a general trend toward  $\alpha$ -selectivity in ether and dioxane, and little (or nonsystematic) selectivity trends in dichloromethane and toluene. Note, however, that when very reactive donors or high temperatures are considered, the glycosylation reaction may become diffusion-controlled, in which case the stereoselectivity may be partly compromised and solvent effects more complicated.<sup>37</sup>

A commonly formulated hypothesis for the mechanism of solvent effects on glycosylation reactions is that a solvent molecule forms a coordination bond with the anomeric carbon of the oxocarbenium ion, preferentially on one side of the ring, thereby blocking the attack by the nucleophile from the same side (Figure 2a).<sup>18</sup> This interpretation will be referred to here as the *solvent coordination hypothesis*. According to this hypothesis, the predominance of the  $\beta$

product (1,2-*trans* glycoside) in acetonitrile would result from the presence of an acetonitrile molecule preferentially coordinated to the anomeric carbon of the oxocarbenium cation on the  $\alpha$  side of the ring. Conversely, the predominance of the  $\alpha$  product (1,2-*cis* glycoside) in ether or dioxane would result from the presence of a solvent molecule preferentially coordinated to the cation from the  $\alpha$  side is apparently supported by experimental investigations of nitrium intermediates in glycosylation reactions.<sup>62</sup> However, these mechanistic studies do not take into account the conformational dynamics of the oxocarbenium cation and the coordination of the counterion. Besides, there is no experimental or theoretical evidence in the literature supporting the suggestion of a preferential coordination of ether or dioxane on the  $\beta$  side.

As will be shown in the present study, quantum-mechanical (QM) calculations on oxocarbenium–solvent interactions in the gas phase and in implicit solvent as well as classical molecular dynamics (MD) simulations of the oxocarbenium intermediate with a triflate counterion (considering a methyl-protected glucopyranoside and the above-mentioned solvents) do not support this hypothesis. Instead, they suggest an alternative mechanism that will be referred to here as the *conformer and counterion distribution hypothesis* (Figure 2b). According to this new hypothesis, the stereoselectivity



**Figure 3.** Chemical structures of the species relevant to the present study. 2,3,4,6-tetra-*O*-methyl-D-glucopyranosyl-triflate (**1**), anomeric isomers of **1** (**1a,b**), oxacarbenium ion (**2**), representative conformers of the oxacarbenium ion (**2a–h**), trifluoromethanesulfonate (triflate) ion (**3**), solvents (**4**, **5**, **6**, **7**), and oxacarbenium-solvent complexes (**8**, **9**), presenting coordination of acetonitrile (**4**) and 1,4-dioxane (**7**) to **2** on the  $\alpha$  side (**8a**, **9a**) or on the  $\beta$  side (**8b**, **9b**).

is explained by solvent-induced variations in the ring conformational preferences of the oxacarbenium cation and in the preferential location of the counterion relative to this cation. Taken together, these effects control the side of the anomeric carbon that can be attacked by the nucleophile. In acetonitrile, the oxacarbenium ion preferentially adopts a B<sub>2,5</sub> boat conformation while the counterion is predominantly located moderately close (on average) to the cation and on the  $\alpha$  side. Both effects prevent the acceptor from attacking from the  $\alpha$  face and enhance the formation of the  $\beta$ -linked product. In contrast, in ether, in toluene, or in dioxane, the oxacarbenium ion preferentially adopts a <sup>4</sup>H<sub>3</sub> half-chair conformation, while the counterion is preferentially located very close to the cation and on the  $\beta$  side. Both effects prevent the acceptor from attacking from the  $\beta$  face and enhance the formation of the  $\alpha$ -linked product. In the present article, the theoretical evidence supporting this alternative conformer and counterion distribution hypothesis as well as the relevance of the alternative solvent coordination hypothesis are described and discussed.

The model system selected for the present theoretical investigations (Figure 3) consists of 2,3,4,6-tetra-*O*-methyl-D-glucopyranosyl-triflate (**1**) as a prototypical glycosyl donor, leading upon leaving group departure to a reactive intermediate complex involving a glucopyranosyl oxacarbenium cation (**2**) and a triflate counterion (**3**). Note that, for simplicity (and

unlike, e.g., the donors considered in Table 1), the leaving group is chosen here to be the same as the counterion. This system was also investigated in previous QM calculations.<sup>68</sup> The conformational properties of the reactive complex are investigated in the solvents acetonitrile (**4**), diethyl ether (**5**), toluene (**6**), and 1,4-dioxane (**7**), as well as in the gas phase (vacuum).

## 2. Computational Methods

**Quantum Mechanical Calculations.** The QM calculations on the reference structures of compounds **1–7** and of the oxacarbenium-solvent complexes **8–9**, as well as on trajectory structures obtained *via* MD simulations (see further below), were all carried out using density functional theory at the B3LYP/6-31G(d,p) level<sup>69,70</sup> in the electronic ground state using the Gaussian 03 program.<sup>71</sup> The calculations on the reference structures **1–9** were performed both in the gas phase (conditions assumed representative for a low polarity solvent such as toluene, dioxane, or ether) and in an implicit solvent (IEF-PCM approach, Integral-Equation Formation-Polarizable Continuum Model<sup>72</sup>) for acetonitrile by using the default parameter of Gaussian 03 for this solvent (dielectric permittivity  $\epsilon = 35.688$ ). Similarly, the trajectory structures obtained *via* MD simulations in dioxane were analyzed in the gas phase, while those obtained *via* MD

simulations in acetonitrile were investigated using the IEF-PCM approach.

The reference structures of **1–7** were generated *via* full geometry optimization in the gas phase for the anomeric isomers of 2,3,4,6-tetra-*O*-methyl-*D*-glucopyranosyl-triflate ( $\alpha$ -anomer **1a** and  $\beta$ -anomer **1b**), initiated from an ideal  ${}^4C_1$  conformation. The energy of **1b** was found to be 28.8 kJ mol $^{-1}$  higher than that of **1a**, as expected from the influence of the anomeric effect.<sup>18,73</sup> The C1–O1 bond lengths in the optimized structures **1a** and **1b** were 0.143 and 0.147 nm, respectively. The energy profile associated with the departure of the triflate anion (**3**) was then calculated by constrained geometry optimization, starting from the optimized structures **1a** and **1b** and progressively elongating the C1–O1 bond from 0.150 to 0.500 nm in steps of 0.005 nm. Removal of the triflate anion (**3**) from the two configurations at maximal elongation led to a unique structure for the oxacarbenium ion (**2**), presenting a  ${}^4H_3$  half-chair conformation (**2a**), in agreement with independent findings.<sup>68</sup> The reference structures of the triflate anion (**3**), as well as of acetonitrile (**4**), diethyl ether (**5**), toluene (**6**), and 1,4-dioxane (**7**) were constructed using the builder function of GaussView<sup>73</sup> followed by full geometry optimization. The reference structures of the corresponding oxacarbenium–solvent complexes (**8a**, **8b**, **9a**, **9b**) were also constructed using GaussView to attach the geometry optimized solvent molecule to the  $\alpha$  or  $\beta$  side of the anomeric carbon of **2a**, followed by full geometry optimization.

Finally, a number of trajectory structures corresponding to the most relevant ring conformations observed during the 100 ns MD simulations of **2** with **3** in solution (see below) were further investigated at the QM level. For these calculations, geometry optimization of the intermediate complex was performed with a constraint on the C1–S distance,  $r$ , to the peak value of the radial distribution function  $P(r)$  obtained from the corresponding MD simulation.

**Molecular Dynamics Simulations.** The explicit-solvent MD simulations were carried out using the GROMOS simulation program<sup>75</sup> together with the 53A6 force field,<sup>76,77</sup> including recently reoptimized parameters for hexopyranose-based carbohydrates.<sup>78–82</sup> Additional parameters required for the description of the oxacarbenium cation were inferred on the basis of the 53A6 glucose molecule (Lennard-Jones and torsional parameters), along with atomic partial charges derived from an electrostatic potential fit based on the QM results. All force-field parameters used in the present study are reported as Supporting Information. The simulations involved an oxacarbenium cation with a triflate counterion in either acetonitrile, ether, toluene, or dioxane. Independent simulations of the pure solvents were performed for the determination of the corresponding dielectric permittivity values.

For the simulations of the solvated oxacarbenium–counterion complexes, the structures of **2** and **3** optimized at the QM level were solvated by 300 solvent molecules at the experimental (room-temperature) density of the solvent<sup>83</sup> (790, 700, 865, 1040 kg m $^{-3}$  for acetonitrile, ether, toluene, and dioxane, respectively), within cubic computational boxes (edge lengths of about 3.0, 3.7, 6.7, and 3.5 nm, respectively).

The pure solvent simulations involved cubic computational boxes containing 1000 solvent molecules at the experimental (room-temperature) density of the solvent.

After energy minimization, initial velocities were assigned from a Maxwell distribution at 300 K, and the systems were equilibrated by 500 ps MD simulation. The production runs were then carried out for a 100 ns (oxacarbenium–counterion complex simulations) or 10 ns (pure solvent simulations) duration. The equations of motion were integrated using a 2 fs time step. The simulations were performed at constant temperature (300 K) and pressure (1 atm) under periodic boundary conditions. The temperature was maintained close to its reference value by weak coupling to a heat bath<sup>84</sup> with a relaxation time of 0.1 ps. The pressure was maintained close to its reference value by weak coupling to a pressure bath<sup>84</sup> (isotropic pressure scaling) with a relaxation time of 0.5 ps and an isothermal compressibility of 0.0004575 (kJ mol $^{-1}$ nm $^{-3}$ ) $^{-1}$ . The center of mass motion was removed every 500 steps. The SHAKE algorithm<sup>84</sup> was applied to constrain the lengths of all covalent bonds and the full rigidity of the solvent molecules. The nonbonded interactions were calculated using a twin-range cutoff scheme,<sup>75,86</sup> with short- and long-range cutoff distances set to 0.8 and 1.4 nm, respectively, and an update frequency of five time steps for the update of the short-range pairlist and intermediate-range interactions. A reaction-field correction was applied to approximately account for the mean effect of electrostatic interactions beyond the long-range cutoff distance. The reaction-field permittivity was set to the corresponding experimental (room temperature) dielectric permittivity of the solvent (38.8, 4.2, 2.4, and 2.2 for acetonitrile,<sup>84</sup> ether,<sup>84</sup> toluene,<sup>87</sup> and dioxane,<sup>88</sup> respectively). Simulations of **2** with **3** in a vacuum at constant temperature (300 K) were also performed for comparison.

The trajectory analyses were performed using the tools of the GROMOS++ software package,<sup>75</sup> as well as several scripts developed for this study. For the oxacarbenium–counterion complex simulations, the conformations of **2** were categorized into eight representative types of conformations of the oxacarbenium ion (Figure 3), namely,  ${}^4H_3$  (**2a**),  ${}^2S_0$  (**2b**),  ${}^{2,5}B$  (**2c**),  ${}^5H_4$  (**2d**),  ${}^3H_4$  (**2e**),  ${}^0S_2$  (**2f**),  $B_{2,5}$  (**2g**), and  ${}^4H_5$  (**2h**). These conformations were defined in terms of the torsional angles  $\gamma_1$ – $\gamma_5$  around five of the six bonds in the pyranose ring, as detailed in Table 2. The selected dihedral angle ranges were slightly adapted from previous definitions,<sup>45,89</sup> so as to permit the attribution of all sampled configurations to one of the recognized ring conformations. The positioning of the triflate anion (**3**) relative to the oxacarbenium cation was characterized by the spherical coordinates  $r$ ,  $\theta$ , and  $\varphi$  describing the direction of the S–C1 vector relative to the local plane of the pyranose ring at C1, as detailed in Figure 4.

For the pure solvent simulations, the permittivity was calculated from the fluctuations of the box dipole moment.

### 3. Results and Discussion

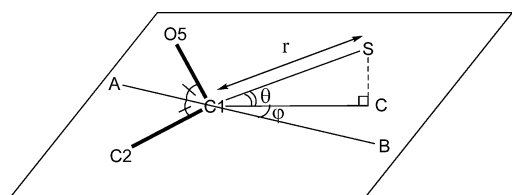
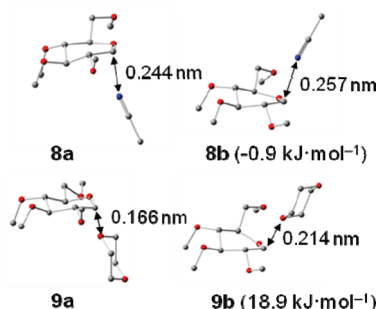
#### QM Calculations on Oxacarbenium-solvent Complexes.

The optimized geometries of the oxacarbenium–solvent complexes **8a**, **8b**, **9a**, and **9b** (Figure 3) are shown in Figure

**Table 2.** Definition of the Dihedral Angles and Corresponding Ranges Used to Assign the Different Ring Conformations of the Oxocarbenium Cation (Figure 3), Namely,  ${}^4\text{H}_3$  (**2a**),  ${}^2\text{S}_0$  (**2b**),  ${}^{2,5}\text{B}$  (**2c**),  ${}^5\text{H}_4$  (**2d**),  ${}^3\text{H}_4$  (**2e**),  ${}^0\text{S}_2$  (**2f**),  $\text{B}_{2,5}$  (**2g**), and  ${}^4\text{H}_5$  (**2h**)

	$\gamma_1$ C5–O5–C1–C2 [deg]	$\gamma_2$ O5–C1–C2–C3 [deg]	$\gamma_3$ C1–C2–C3–C4 [deg]	$\gamma_4$ C2–C3–C4–C5 [deg]	$\gamma_5$ C3–C4–C5–O5 [deg]
<b>2a</b>	0 ± 60	0 ± 80	310 ± 30	50 ± 30	310 ± 30
<b>2b</b>	0 ± 60	0 ± 80	310 ± 30	50 ± 30	0 ± 20
<b>2c</b>	0 ± 60	0 ± 80	310 ± 30	0 ± 20	50 ± 30
<b>2d</b>	0 ± 60	0 ± 80	359 ± 19	310 ± 30	50 ± 30
<b>2e</b>	0 ± 60	0 ± 80	49 ± 31	310 ± 30	50 ± 30
<b>2f</b>	0 ± 60	0 ± 80	49 ± 31	310 ± 30	350 ± 30
<b>2g</b>	0 ± 60	0 ± 80	49 ± 31	5 ± 25	320 ± 40
<b>2h</b>	0 ± 60	0 ± 80	359 ± 19	49 ± 31	320 ± 40

5. The potential energy of the  $\beta$  complex (**8b**) is found to be slightly lower than that of the  $\alpha$  complex (**8a**) for acetonitrile (by 0.9 kJ mol $^{-1}$  using the IEF-PCM solvation model; 4.23 kJ mol $^{-1}$  in the gas phase), while the potential energy of the  $\alpha$  complex (**9a**) is found to be much lower than that of the  $\beta$  complex (**9b**) for dioxane (gas phase calculation). The above observations are clearly at odds with the *solvent coordination hypothesis* (Figure 2a). This hypothesis would imply a preferential coordination of acetonitrile on the  $\alpha$  side (favoring a nucleophilic attack from the  $\beta$  side and leading to the experimentally observed predominance of the  $\beta$  product) and a preferential coordina-

**Figure 4.** Definition of the spherical coordinates  $r$ ,  $\theta$ , and  $\varphi$  used to characterize the positioning of the triflate anion relative to the oxocarbenium cation. These coordinates describe the direction of the C1–S vector relative to the local plane of the pyranose ring at C1. Line AB is the interior bisector of the angle C2–C1–O5. Line SC is perpendicular to the plane C2–C1–O5, C being the intersection point. The distance  $r$  corresponds to the length of the C1–S vector. The angle  $\theta$  corresponds to the angle C–C1–S. The angle  $\varphi$  corresponds to the angle B–C1–C.**Figure 5.** Geometry-optimized structures (QM) of the oxocarbenium-solvent complexes **8** and **9** (Figure 3) involving a coordinated acetonitrile molecule (calculation using an IEF-PCM implicit solvent model) or dioxane molecule (calculation in the gas phase). The distance between the anomeric carbon and the nitrogen atom for **8** or the oxygen atom for **9**, as well as the potential energies of the  $\beta$  complexes (**8b**, **9b**) relative to the  $\alpha$  complexes (**8a**, **9a**), are also indicated.**Table 3.** Ratios of Ring Conformers (Figure 3) of the Oxocarbenium Ion Observed during the 100 ns MD Simulations of the Oxocarbenium–Counterion Complex in the Different Solvents (As Well As in Vacuum)

solvent	ratio of conformers [%]							
	<b>2a</b>	<b>2b</b>	<b>2c</b>	<b>2d</b>	<b>2e</b>	<b>2f</b>	<b>2g</b>	<b>2h</b>
<b>4</b>				0.0	36.5	60.7	2.7	0.0
<b>5</b>				0.0	50.1	47.6	2.3	0.0
<b>6</b>			0.0	0.0	70.9	27.7	1.3	0.0
<b>7</b>				0.0	61.4	37.0	1.6	0.0
in vacuum				0.1	68.8	29.6	1.5	0.0

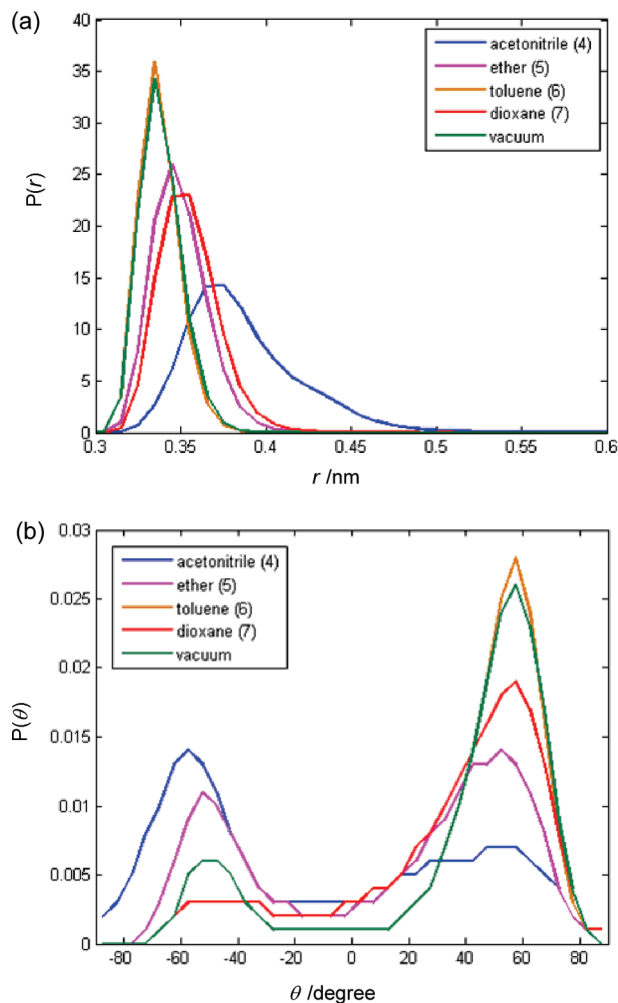
tion of dioxane on the  $\beta$  side (favoring a nucleophilic attack from the  $\alpha$  side and leading to the experimentally observed predominance of the  $\alpha$  product), i.e., a trend exactly opposite to that suggested by the present QM calculations.

**MD Simulations of the Solvated Oxocarbenium–Counterion Complex.** As a preliminary calculation, the dielectric permittivities of the solvent models employed for acetonitrile (**4**), ether (**5**), toluene (**6**), and dioxane (**7**) were calculated on the basis of pure solvent MD simulation and were found to be 34.6, 3.5, 1.0, and 1.1, respectively, in good qualitative agreement with the corresponding (room-temperature) experimental values of 35.8, 4.3, 2.4, and 2.2.<sup>83,87,88</sup>

The populations of ring conformers (Figure 3) of the oxocarbenium ion observed during the MD simulations of the oxocarbenium–counterion complexes in the different solvents as well as in vacuum are reported in Table 3. Although all simulations were initiated from the same  ${}^4\text{H}_3$  conformation (**2a**), the equilibrium distribution encompasses  ${}^{2,5}\text{B}$  (**2c**),  ${}^5\text{H}_4$  (**2d**),  ${}^3\text{H}_4$  (**2e**),  ${}^0\text{S}_2$  (**2f**),  $\text{B}_{2,5}$  (**2g**), and  ${}^4\text{H}_5$  (**2h**) conformers. The  ${}^3\text{H}_4$  (**2e**) and the  ${}^0\text{S}_2$  (**2f**) conformers are dominant in all simulations, the proportion of the latter increasing with the polarity of the solvent. Thus, for example, the oxocarbenium ion preferentially adopts a  ${}^0\text{S}_2$  (**2f**) conformation rather than a  ${}^3\text{H}_4$  (**2e**) conformation in acetonitrile (**2e/2f** ratio of 36.5:60.7), whereas the opposite is observed in dioxane (**2e/2f** ratio of 61.4:37.0). The proportions observed in toluene or in vacuum are close to those found in dioxane, while the proportions observed in ether are intermediate between those found in acetonitrile and dioxane.

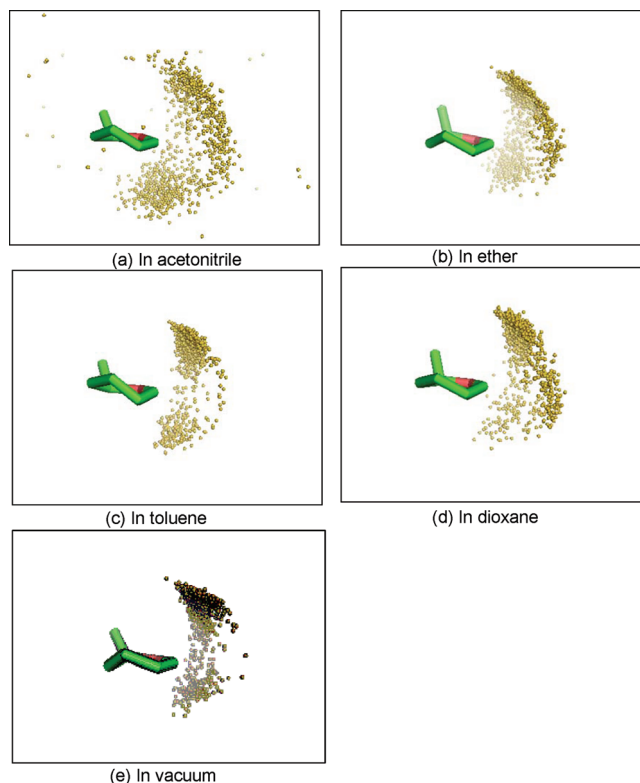
The results of the MD simulations concerning the positioning of the counterion relative to the oxocarbenium cation in the different solvents are illustrated graphically in Figures 6 and 7, and summarized numerically in Table 4. The following observations can be made. First, the probability





**Figure 6.** Positioning of the counterion relative to the oxocarbenium cation (Figure 4) observed during the 100 ns MD simulations of the oxocarbenium–counterion complex in the different solvents (as well as in vacuum): (a) distribution  $P(r)$  of the distance  $r$ , (b) distribution  $P(\theta)$  of the angle  $\theta$ . The  $P(\theta)$  functions are normalized according to  $\int_{-\pi/2}^{\pi/2} P(\theta) \cos(\theta) d\theta = 1$ . The probability  $P(\theta)$  was also calculated for acetonitrile with a cutoff  $r \leq 0.375$  nm, the distance value at the peak of  $P(r)$ , resulting in a nearly identical distribution (not shown).

distribution  $P(r)$  of the C1–S distance  $r$  (Figure 6a) tends to become broader, i.e., stretched to larger distances, with an increase of the polarity of the solvent. Second, the probability distribution  $P(\theta)$  of the angle  $\theta$  formed by the C1–S vector and the local ring plane at C1 (Figure 6b) is bimodal, with peaks at about  $\pm 55^\circ$  ( $\theta < 0^\circ$ ,  $\alpha$  side;  $\theta > 0^\circ$ ,  $\beta$  side). The population associated with the  $\theta = -55^\circ$  peak increases (relative to that associated with the  $\theta = +55^\circ$  peak) with increasing polarity of the solvent. In other words, the simulation results show that an increase in the solvent polarity leads to a less tight binding of the triflate counterion to the oxocarbenium cation and to a progressive shift of its preferential positioning from the  $\beta$  to the  $\alpha$  side of the ring. This trend is also clearly evident when considering the distributions of the counterion (successive positions of the triflate S atom) along the different trajectories (Figure 7). The correlation between the breadth of the cation–anion distance distribution and the solvent polarity is easily



**Figure 7.** Positioning of the counterion relative to the oxocarbenium cation observed during the 100 ns MD simulations of the oxocarbenium–counterion complex in the different solvents (as well as in vacuum). Successive positions of the triflate S atom along the trajectories are displayed at 100 ps intervals (1000 yellow beads), after superimposition of the trajectory frames onto the initial configuration ( ${}^4\text{H}_3$  ring conformation of the oxocarbenium ion) based on all carbon atoms of the cation.

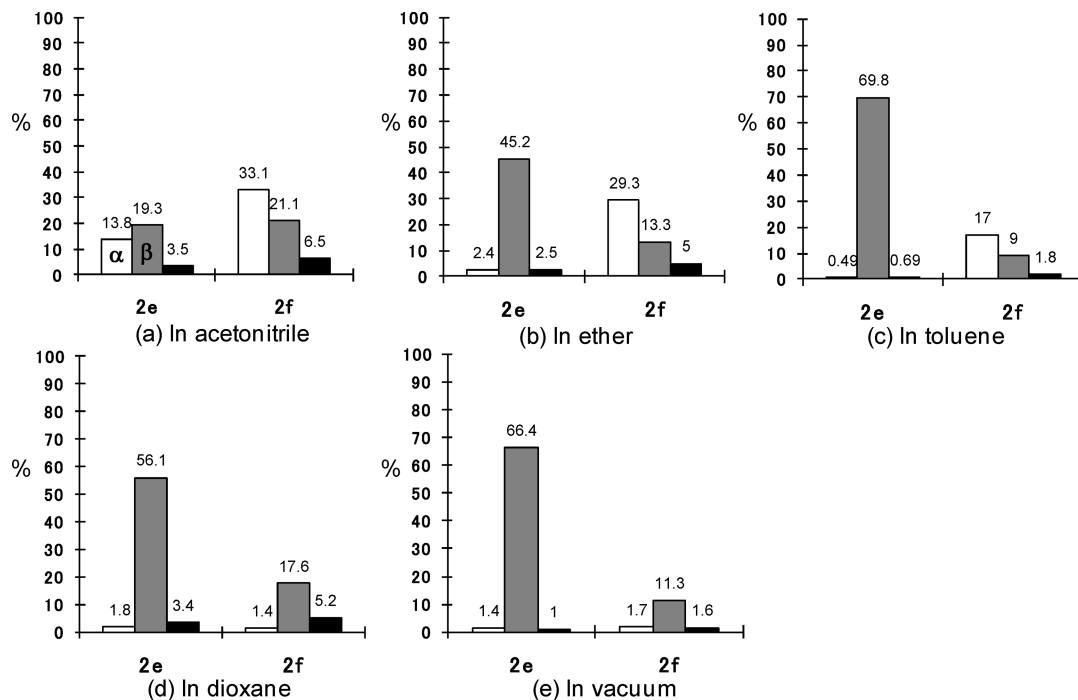
**Table 4.** Positioning of the Counterion Relative to the Oxocarbenium Cation (Figures 4 and 6) Observed during the 100 ns MD Simulations of the Oxocarbenium–Counterion Complex in the Different Solvents (As Well As in Vacuum)

solvent	$\theta \leq -10^\circ$ ( $\alpha$ -side) [%]	$\theta \geq 10^\circ$ ( $\beta$ -side) [%]	$-10^\circ < \theta < 10^\circ$ [%]	$\tau$ [nm]	$ \bar{\varphi} $ [deg]
4	48.2	41.6	10.2	0.438	53.0
5	33.5	58.9	75.6	0.349	39.8
6	18.4	79.1	2.5	0.338	32.3
7	16.9	74.3	8.8	0.355	35.8
in vacuum	19.2	78.2	2.7	0.338	32.7

explained in terms of dielectric screening effects (counteracting the direct Coulombic attraction between the two ions). The concomitant shift from a preferential  $\beta$ -side to a preferential  $\alpha$ -side coordination upon increasing the solvent polarity is more difficult to rationalize and appears to be correlated with the shift from a dominant  ${}^3\text{H}_4$  (**2e**) to a dominant  ${}^0\text{S}_2$  (**2f**) conformation (see further below).

As a result of these effects, in acetonitrile, the anion presents a slight preference for the  $\alpha$  side ( $\theta < -10^\circ$ ) compared to the  $\beta$  side ( $\theta > 10^\circ$ ), with an  $\alpha/\beta$  ratio of 48.2:41.6 (Table 4). In contrast, it is predominantly found on the  $\beta$  side in ether (33.5:58.9), toluene (18.4:79.1), dioxane (16.9:74.3), and in vacuum (19.2:78.2). The corresponding average





**Figure 8.** Correlation between the ring conformation of the oxocarbenium cation and the preferential counterion positioning, as observed during the 100 ns MD simulations of the oxocarbenium–counterion complex in the different solvents (as well as in vacuum). The proportions of the sampled configurations corresponding to  ${}^3\text{H}_4$  (**2e**) and  ${}^0\text{S}_2$  (**2f**) ring conformers, along with  $\alpha$ -side ( $\theta \leq -10^\circ$ ; white bars),  $\beta$ -side ( $\theta \geq 10^\circ$ ; gray bars), or in-plane ( $-10^\circ < \theta < 10^\circ$ ; black bars) counterion locations, are reported.

values of  $r$  and  $|\varphi|$  also show that the triflate ion is more weakly bound to the anomeric carbon in acetonitrile ( $\bar{r} = 0.438$  nm,  $|\bar{\varphi}| = 53.0^\circ$ ) compared to the other solvents ( $\bar{r}$  ranging from 0.338 to 0.355 nm, and  $|\bar{\varphi}|$  ranging from 32.3 to 39.8°).

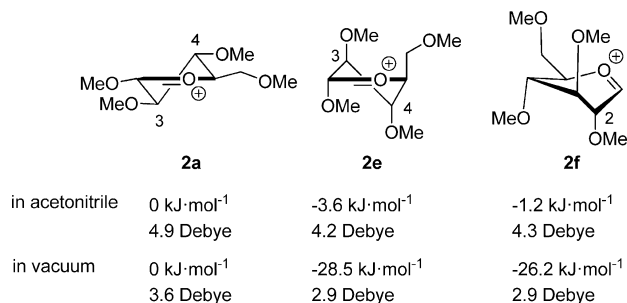
In order to further investigate the correlation between the ring conformation of the oxocarbenium ion and the preferential counterion positioning, the latter positioning was analyzed separately for the configurations presenting  ${}^3\text{H}_4$  (**2e**) or  ${}^0\text{S}_2$  (**2f**) conformations of the pyranose ring in the different solvents. In acetonitrile (Figure 8a), the dominant configurations involve a  ${}^0\text{S}_2$  (**2f**) ring conformation with the counterion on the  $\alpha$  side (**2f- $\alpha$** ; 33.1%), while three alternative configurations are nearly equally populated, namely, **2f- $\beta$**  (21.1%), **2e- $\beta$**  (19.3%), and **2e- $\alpha$**  (13.8%). In the other solvents (Figure 8b–e), the oxocarbenium ion predominantly adopts a  ${}^3\text{H}_4$  conformation (**2e**) with the counterion on the  $\beta$  side (**2e- $\beta$** ; 45.2%, 69.8%, 56.1%, and 66.4% in ether, toluene, dioxane, and a vacuum, respectively). In dioxane and a vacuum, there is almost no oxocarbenium ion with the counterion at its  $\alpha$  side, and the configurations **2e- $\beta$**  and **2f- $\beta$**  taken together account for 73.7 and 77.7%, respectively, of the all sampled configurations.

In summary, the MD simulations suggest that, in acetonitrile, the oxocarbenium ion preferentially adopts a  ${}^0\text{S}_2$  (**2f**) or, to a lesser extent, a  ${}^3\text{H}_4$  (**2e**) conformation, with the counterion loosely bound to the anomeric carbon and distributed nearly equally on the  $\alpha$  and  $\beta$  sides (slight  $\alpha$  side preference for **2f** and  $\beta$  side preference for **2e**). In contrast, in the solvents of lower polarity, the oxocarbenium ion preferentially adopts a  ${}^3\text{H}_4$  (**2e**) conformation with the

counterion tightly bound to the anomeric carbon and predominantly on the  $\beta$  side. This  $\beta$  side preference is also observed for the minor  ${}^0\text{S}_2$  conformer.

The above observations provide support to the *conformer and counterion distribution hypothesis* (Figure 2b). This hypothesis suggests that, in acetonitrile, the preferential ring conformation ( ${}^0\text{S}_2$ ) and counterion positioning ( $\alpha$  side) both favor an attack of the nucleophile from the  $\beta$  side, leading to the experimentally observed predominance of the  $\beta$  product, while in a solvent of lower polarity, the preferential ring conformation ( ${}^3\text{H}_4$ ) and counterion positioning ( $\beta$  side) both favor an attack of the nucleophile from the  $\alpha$  side (leading to the experimentally observed predominance of the  $\alpha$  product). This is in excellent agreement with the results of the present MD simulations for acetonitrile, ether, and dioxane. Note that this interpretation differs from the previously formulated hypothesis<sup>49</sup> that  ${}^4\text{H}_3$  conformers are preferentially  $\alpha$ -selective and  ${}^3\text{H}_4$  conformers  $\beta$ -selective, indicating that the consideration of the counterion positioning is essential in the theoretical investigation of glycosylation intermediates.

The simulation results are, however, in apparent contradiction with the *conformer and counterion distribution hypothesis* in the case of toluene. For this solvent, the conformational properties of the oxocarbenium–counterion complex are similar to those observed in ether, dioxane, and vacuum. However, synthetic experiments usually observe no or a low stereoselectivity in toluene.<sup>60,61,63–65,67</sup> One possible reason for this discrepancy is that the conformational properties of the oxocarbenium–counterion complex in toluene are influenced by effects that are not



**Figure 9.** Geometry-optimized structures (QM) of the oxacarbenium cation (**2**) in <sup>4</sup>H<sub>3</sub> (**2a**), <sup>3</sup>H<sub>4</sub> (**2e**), and <sup>0</sup>S<sub>2</sub> (**2f**) ring conformations in acetonitrile (calculation using an IEF-PCM implicit solvent model) or in the gas phase (assumed representative for dioxane). The potential energies (relative to **2a**) and the dipole moments (relative to the center of charge) are also indicated.

taken into account appropriately in classical force-field simulations, such as stereoelectronic effects and cation– $\pi$  interactions. Note, however, that a significant (nonsystematic) stereoselectivity may also be observed in this solvent, depending on the donor, acceptor, and activator (e.g., entries 7 and 12 in Table 1, evidencing clear  $\alpha$ - and  $\beta$ -stereoselectivities, respectively).

#### QM Analysis of Selected MD Trajectory Structures.

The classical force-field representation employed in the MD simulations takes realistically into account solvation effects but has its shortcomings, including an approximate description of stereoelectronic effects (controlling in particular the relative stabilities of the different ring conformations). For this reason, selected configurations of the oxacarbenium–counterion complex presenting different ring conformations were extracted from the simulations in acetonitrile and dioxane and subjected to a QM analysis, i.e., geometry optimization and energy evaluation. These calculations were performed using the IEF-PCM implicit solvent model (acetonitrile) or in the gas phase (dioxane).

In a first step, configurations of the oxacarbenium ion without the counterion were geometry optimized starting from MD configurations presenting the <sup>3</sup>H<sub>4</sub> (**2e**) or the <sup>0</sup>S<sub>2</sub> (**2f**) conformations. This optimization did not result in ring

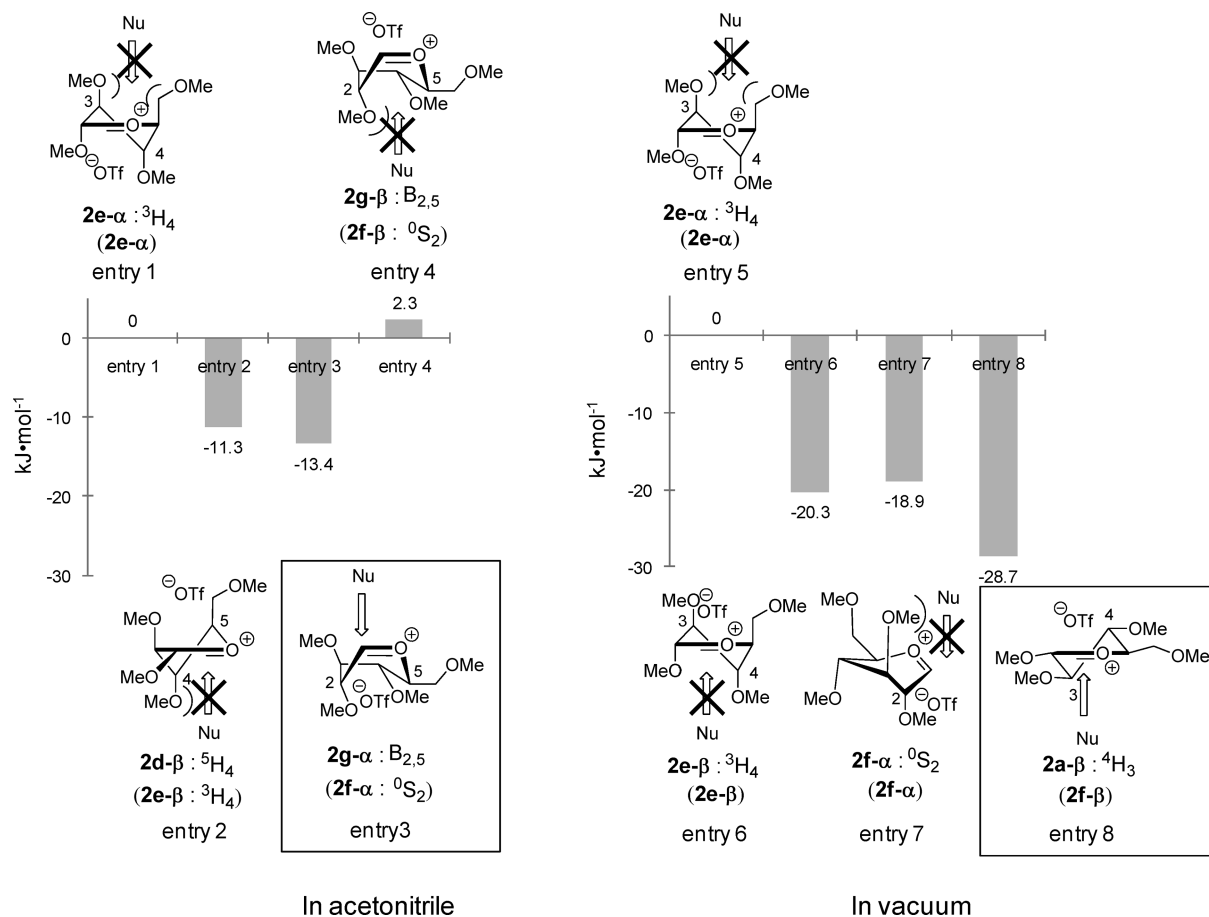
conformational changes (for <sup>3</sup>H<sub>4</sub>, this observation contradicts previous suggestions concerning the stability of this conformer during QM calculations<sup>55</sup>). The final energies and dipole moments, in both acetonitrile and in vacuum (dioxane), are reported in Figure 9 and compared to the corresponding values for the geometry optimized <sup>4</sup>H<sub>3</sub> conformation (**2a**). The latter conformation is often considered to be the most stable one for a typical oxacarbenium cation<sup>36,89–92</sup> and was used as the starting conformation for the MD simulations. Both of the structures predominantly sampled during the MD simulations (**2e**, **2f**) are about 1–3 and 26–28 kJ mol<sup>-1</sup> more stable than **2a** in acetonitrile and in vacuum, respectively. The energy ranking of the three structures matches that of the dipole moment magnitudes; i.e., the favored conformations (**2e** and, to a slightly lesser extent, **2f**) are those with the lowest dipole moment. Expectedly, the energy differences are much larger in vacuum (dioxane) compared to acetonitrile, because a polar medium more efficiently stabilizes conformations with a higher dipole moment. The observed trends are in good qualitative agreement with the relationships inferred from the MD simulations between preferential ring conformations and solvent polarity.

In a second step, configurations of the oxacarbenium–counterion complex were geometry optimized starting from MD configurations presenting a <sup>3</sup>H<sub>4</sub> conformation with the counterion on the  $\alpha$  side (**2e- $\alpha$** ) or  $\beta$  side (**2e- $\beta$** ), or a <sup>0</sup>S<sub>2</sub> conformation with the counterion on the  $\alpha$  side (**2f- $\alpha$** ) or  $\beta$  side (**2f- $\beta$** ). These optimizations were performed with a constraint on the C1–S distance, set to the peak value of  $P(r)$  (Figure 6a) as determined in the MD simulations in acetonitrile (0.38 nm) or in dioxane (0.34 nm). In four cases, the optimization resulted in a conformational change of the ring to either a <sup>3</sup>H<sub>4</sub> (**2d**), a B<sub>2,5</sub> (**2g**), or a <sup>4</sup>H<sub>3</sub> (**2a**) configuration, namely, **2e- $\beta$**   $\rightarrow$  **2d- $\beta$** , **2f- $\alpha$**   $\rightarrow$  **2g- $\alpha$** , and **2f- $\beta$**   $\rightarrow$  **2g- $\beta$**  in acetonitrile and **2f- $\beta$**   $\rightarrow$  **2a- $\beta$**  in vacuum (dioxane). The final energies and dipole moments, in both acetonitrile and vacuum (dioxane), are reported in Table 5 and compared to the relative conformer populations (**2e- $\alpha$** , **2e- $\beta$** , **2f- $\alpha$** , and **2f- $\beta$** ) observed in the corresponding MD simulations (Figure 8).

**Table 5.** Relative Energies Corresponding to Geometry-Optimized Structures (QM) of the Oxacarbenium Cation (**2**) Presenting Different Ring Conformations (Figure 3) and Counterion (**3**) Positioning ( $\alpha$  or  $\beta$ ) in Acetonitrile (Calculation Using an IEF-PCM Implicit Solvent Model) or in Vacuum (Assumed Representative for Dioxane)<sup>a</sup>

entry no.	starting conformation of <b>2</b>	location of <b>3</b>	ratio in MD simulations [%]	conformation of <b>2</b> in optimized complex	relative energy [kJ mol <sup>-1</sup> ]	dipole moment [Debye]
in acetonitrile						
1	<b>2e</b>	$\alpha$	13.8	<b>2e</b>	0	19.2
2		$\beta$	19.3	<b>2d</b>	-11.3	15.6
3	<b>2f</b>	$\alpha$	33.1	<b>2g</b>	-13.4	17.8
4		$\beta$	21.1	<b>2g</b>	2.3	17.9
in vacuum						
5	<b>2e</b>	$\alpha$	0.17	<b>2e</b>	0	10.0
6		$\beta$	56.1	<b>2e</b>	-20.3	9.5
7	<b>2f</b>	$\alpha$	1.4	<b>2f</b>	-18.9	10.7
8		$\beta$	17.6	<b>2a</b>	-28.7	9.7

<sup>a</sup> The energies are given relative to entries 1 (in acetonitrile) or 5 (in vacuum). The corresponding dipole moments (relative to the center of charge) are also reported. The initial ring conformation prior to geometry optimization and the occurrence of this specific configuration in the MD simulations are also indicated.



**Figure 10.** Relative potential energies corresponding to geometry-optimized structures (QM) of the oxocarbenium cation (**2**) presenting different ring conformations and counterion (**3**) positioning ( $\alpha$  or  $\beta$  side) in acetonitrile (calculation using an IEF-PCM implicit solvent model) or in vacuum (assumed representative for dioxane). The entry numbers refer to Table 5. The energies are given relative to entries 1 (in acetonitrile) or 5 (in vacuum). The initial ring conformation prior to geometry optimization is indicated between parentheses.

Although the MD simulations and QM calculations suggest slightly different dominant configurations for the reactive oxocarbenium–counterion complex in acetonitrile (**2f-α** and **2g-α**, respectively), these configurations present two common features. First, the coordination of the counterion on the  $\alpha$  face blocks a nucleophilic attack from this side. Second, the lack of steric crowding and high exposure of the anomeric carbon toward the  $\beta$  face facilitates a nucleophilic attack from this side (Figure 10). In contrast, **2e-α**, **2d-β**, and **2g-β** are sterically crowded on the face opposite the counterion. These observations are consistent with the experimentally observed predominance of the  $\beta$ -product in acetonitrile. Note that the conformations identified here do not include the <sup>5</sup>S<sub>1</sub> conformer, which was proposed previously on the basis of QM calculations on the same compound in dichloromethane,<sup>91</sup> this conformation being similar to **2g** except for the pseudoequatorial orientation of C2–O2 bond.

The MD simulations and QM calculations also suggest different dominant configurations for the reactive oxocarbenium–cation complex in dioxane (**2e-β** and **2a-β**, respectively). However these configurations again present two common features. First, the coordination of the counterion on the  $\beta$  face blocks a nucleophilic attack from this side.

Second, the lack of steric crowding and high exposure of the anomeric carbon toward the  $\alpha$  face facilitates a nucleophilic attack from this side (Figure 10; although the exposure is similar, the steric crowding is slightly higher for **2e-β** compared to **2a-β**). In contrast, **2e-α** and **2f-α** are sterically crowded on the face opposite the counterion. These observations are consistent with the experimentally observed predominance of the  $\alpha$  product in dioxane.

The suggestion of a <sup>4</sup>H<sub>3</sub> conformation with the counterion on the  $\beta$  side (**2a-β**) for the reactive intermediate complex in solvents of low polarity ( $\alpha$ -selectivity) is also compatible with the result of glycosylation experiments involving conformationally locked pyranosides functionalized by a *N*-benzyl-2,3-*trans*-oxazolidinone group.<sup>93–96</sup> For these compounds, <sup>4</sup>H<sub>3</sub> is the only possible conformation of the pyranose ring, and in agreement with the present suggestion, these compounds predominantly lead to the formation of  $\alpha$ -linked glycosides.

To our knowledge, the present work is the first study suggesting a key role for the (solvent-modulated) counterion coordination in influencing the stereoselectivity of glycosylation reactions, besides a previously postulated role of this anion as a proton acceptor near the transition state of the reaction.<sup>55</sup>

## 4. Conclusions

The present study combines QM calculations and explicit-solvent MD simulations to gain a better understanding of solvent effects on the stereoselectivity of glycosylation reactions. To this purpose, a model system consisting of a methyl-protected triflate glucopyranoside in different solvents (acetonitrile, ether, dioxane, toluene, and in vacuum) is considered.

The common assumption concerning solvent effects on the stereoselectivity of glycosylation reactions, the *solvent coordination hypothesis*, suggests that the preferential coordination of a solvent molecule to the reactive oxacarbenium cation on one side of the anomeric carbon ( $\alpha$  or  $\beta$ ) hinders a nucleophilic attack from this side, thereby favoring the product with the opposite stereochemistry ( $\beta$  or  $\alpha$ ). The present calculations do not support this hypothesis. For example, an acetonitrile molecule is predicted to preferentially bind on the  $\beta$  side, in disagreement with the experimentally observed  $\alpha$ -selectivity in this solvent. Conversely, a dioxane molecule is predicted to preferentially bind on the  $\alpha$  side, in disagreement with the experimentally observed  $\alpha$ -selectivity.

However, the present calculations support an alternative explanation, termed here the *conformer and counterion distribution hypothesis*. This new hypothesis suggests that the stereoselectivity is dictated by two interrelated conformational properties of the reactive oxacarbenium-counterion complex, namely, (1) the conformational preferences of the oxacarbenium pyranose ring, modulating the steric crowding and exposure of the anomeric carbon toward the  $\alpha$  or  $\beta$  face, and (2) the preferential coordination of the counterion to the oxacarbenium cation on one side of the anomeric carbon, hindering a nucleophilic attack from this side. For example, in acetonitrile, the calculations suggest a dominant  ${}^0S_2$  (MD) or  $B_{2,5}$  (QM) ring conformation of the oxacarbenium ion with preferential coordination of the counterion on the  $\alpha$  side within the reactive intermediate complex. Both factors render the anomeric carbon most accessible from the  $\beta$  side, in agreement with the experimentally observed  $\beta$ -selectivity in this solvent. Conversely, in dioxane, the calculations suggest a dominant  ${}^3H_4$  (MD) or  ${}^4H_3$  (QM) ring conformation with preferential counterion coordination on the  $\beta$  side. Both factors render the anomeric carbon most accessible from the  $\alpha$  side, in agreement with the experimentally observed  $\alpha$  selectivity in this solvent. The reactive conformations predicted by the QM calculations ( $B_{2,5}$  in acetonitrile and  ${}^4H_3$  in dioxane) are probably more realistic, since these calculations take more appropriately into account the stereoelectronic effects controlling the relative stabilities of the different ring conformations.

In the case of dioxane, the  ${}^4H_3$  conformation is indeed the one usually considered to be the most stable for typical oxacarbenium cations.<sup>36,90–92</sup> It is also the only possible conformation in the case of conformationally locked pyranosides functionalized by a *N*-benzyl-2,3-*trans*-oxazolidinone group, which predominantly lead to  $\alpha$ -linked disaccharides upon glycosylation.<sup>93–96</sup> In the case of acetonitrile, the suggestion of a reactive  $B_{2,5}$  conformation has been formulated previously on the basis of experiments on *N*-(tetra-O-

acetyl- $\alpha$ -D-glucopyranosyl-4-methyl-pyridinium bromide) in aqueous solution.<sup>97,98</sup> The shift between the former and latter conformations upon increasing the polarity of the solvent may tentatively be attributed to dielectric screening effects, increasingly stabilizing conformers with higher dipole moments.

In summary, the theoretical (and experimental) data discussed in the present study are clearly compatible with the new *conformer and counterion distribution hypothesis* and do not provide support to the more common *solvent coordination hypothesis*. However, mechanistic hypotheses can seldom be formally “proved”. They can only be strengthened by accumulation of compatible data and elimination of concurrent hypotheses. In this sense, the present work provides preliminary evidence for a key role of the oxacarbenium conformation and counterion coordination within the reactive complex. Theoretical and experimental work is currently in progress to further refine this new hypothesis.

**Acknowledgment.** The present calculations were performed on Obelix, a computer cluster of the Competence Center for Computational Chemistry (C<sup>4</sup>). The authors would like to thank Hans Peter Lüthi, Maria Reif, and the members of the group for computer-aided chemistry (IGC) for their help and for many valuable discussions.

**Supporting Information Available:** All force-field parameters used in the present study as well as the experimental procedures employed and the  ${}^1H$  NMR and  ${}^{13}C$  NMR spectral characteristics of the two disaccharides ( $\alpha$ - and  $\beta$ -linked) newly synthesized for the present study (entries 9–14 of Table 1) are provided. This information is available free of charge *via* the Internet at <http://pubs.acs.org/>.

## References

- (1) Stallforth, P.; Lepenies, B.; Adibekian, A.; Seeberger, P. H. Carbohydrates: A Frontier in Medicinal Chemistry. *J. Med. Chem.* **2009**, *52*, 5561–5577.
- (2) Seeberger, P. H. Chemical glycobiology: why now. *Nature Chem. Biol.* **2009**, *5*, 368–372.
- (3) The stereoselectivity is often described using relative stereo-descriptors to describe the configuration at the anomeric carbon,  $\alpha$  and  $\beta$ , instead of the notation 1,2-*cis* and 1,2-*trans*. The  $\alpha$  descriptor is used for the case where the OH at the anomeric carbon is on the same side as the OH at the reference atom in the Fischer projection, and the  $\beta$  descriptor is used for the opposite side. In the case of glucopyranosides, the 1,2-*cis* and 1,2-*trans* positions are defined as the  $\alpha$  and  $\beta$  configurations for the anomeric carbon, respectively. For more details, see the IUPAC definition at <http://www.chem.qmul.ac.uk/iupac/2carb/06n07.html> (accessed April 13, 2010).
- (4) Manabe, S.; Ito, Y. On-Resin Real-Time Reaction Monitoring of Solid-Phase Oligosaccharide Synthesis. *J. Am. Chem. Soc.* **2002**, *124*, 12638–12639.
- (5) Simon, J.; Liu, K.; Schmidt, R. R. Solid-Phase Oligosaccharide Synthesis of a Small Library of *N*-Glycans. *Chem.—Eur. J.* **2006**, *12*, 1274–1290.
- (6) Ando, H.; Manabe, S.; Nakahara, Y.; Ito, Y. Tag-Reporter Strategy for Facile Oligosaccharide Synthesis on Polymer Support. *J. Am. Chem. Soc.* **2001**, *123*, 3848–3849.



- (7) Ando, H.; Manabe, S.; Nakahara, Y.; Ito, Y. Solid-Phase Capture-Release Strategy Applied to Oligosaccharide Synthesis on a Soluble Polymer Support. *Angew. Chem., Int. Ed.* **2001**, *40*, 4725–4728.
- (8) Hanashima, S.; Manabe, S.; Ito, Y. Divergent Synthesis of Sialylated Glycan Chains: Combined Use of Polymer Support, Resin Capture-Release, and Chemoenzymatic Strategies. *Angew. Chem., Int. Ed.* **2005**, *44*, 4218–4224.
- (9) Kantchev, E. A. B.; Bader, S. J.; Parquette, J. R. Oligosaccharide Synthesis on a Soluble, Hyperbranched Polymer Support Via Thioglycoside Activation. *Tetrahedron* **2005**, *61*, 8329–8338.
- (10) Bauer, J.; Rademann, J. Hydrophobically Assisted Switching Phase Synthesis: The Flexible Combination of Solid-Phase and Solution-Phase Reactions Employed for Oligosaccharide Preparation. *J. Am. Chem. Soc.* **2005**, *127*, 7296–7297.
- (11) Fukase, K.; Takashina, M.; Hori, Y.; Tanaka, D.; Tanaka, K.; Kusumoto, S. Oligosaccharide Synthesis by Affinity Separation Based on Molecular Recognition between Podand Ether and Ammonium Ion. *Synlett* **2005**, 2342–2346.
- (12) Goto, K.; Miura, T.; Hosaka, D.; Matsumoto, H.; Mizuno, M.; Ishida, H.; Inazu, T. Rapid Oligosaccharide Synthesis on a Fluorous Support. *Tetrahedron* **2004**, *60*, 8845–8854.
- (13) Barresi, F.; Hindsgaul, O. In *Modern Methods in Carbohydrate Synthesis*; Khan, S. H., O'Neil, R. A., Eds.; Harwood Academic Publishers: Amsterdam, 1996; pp 251–276.
- (14) Miliković, M.; Yeagley, D.; Deslongchamps, P.; Dory, Y. L. Experimental and Theoretical Evidence of Through-Space Electrostatic Stabilization of the Incipient Oxocarbenium Ion by an Axially Oriented Electronegative Substituent During Glycopyranoside Acetolysis. *J. Org. Chem.* **1997**, *62*, 7597–7604.
- (15) Crich, D.; Sun, S. Are Glycosyl Triflates Intermediates in the Sulfoxide Glycosylation Method? A Chemical and  $^1\text{H}$ ,  $^{13}\text{C}$ , and  $^{19}\text{F}$  NMR Spectroscopic Investigation. *J. Am. Chem. Soc.* **1997**, *119*, 11217–11223.
- (16) Pozsgay, V. In *Carbohydrates in Chemistry and Biology*; Ernst, B., Hart, G. W., Sinay, P., Eds.; Wiley-VCH: Weinheim, Germany, 2000.
- (17) Crich, D. Chemistry of Glycosyl Triflates: Synthesis of  $\beta$ -Mannopyranosides. *J. Carbohydr. Chem.* **2002**, *21*, 663–686.
- (18) Demchenko, A. V. Stereoselective Chemical 1,2-cis O-Glycosylation: From 'Sugar Ray' to Modern Techniques of the 21st Century. *Synlett* **2003**, 1225–1240.
- (19) Crich, D.; Chandrasekara, N. S. Mechanism of 4,6-O-Benzylidene-Directed  $\beta$ -Mannosylation as Determined by  $\alpha$ -Deuterium Kinetic Isotope Effects. *Angew. Chem., Int. Ed.* **2004**, *43*, 5386–5389.
- (20) Yamago, S.; Yamada, T.; Maruyama, T.; Yoshida, J. Iterative Glycosylation of 2-Deoxy-2-aminothioglycosides and Its Application to the Combinatorial Synthesis of Linear Oligoglucosamines. *Angew. Chem., Int. Ed.* **2004**, *43*, 2145–2148.
- (21) Wei, P.; Kerns, R. J. Factors Affecting Stereocontrol during Glycosidation of 2,3-Oxazolidinone-Protected 1-Tolythio-N-acetyl-D-glucosamine. *J. Org. Chem.* **2005**, *70*, 4195–4198.
- (22) Horenstein, N. A. Mechanisms for Nucleophilic Aliphatic Substitution at Glycosides. *Adv. Phys. Org. Chem.* **2006**, *41*, 275–314.
- (23) Rencurosi, A.; Lay, L.; Russo, G.; Caneva, E.; Poletti, L. NMR Evidence for the Participation of Triflated Ionic Liquids in Glycosylation Reaction Mechanisms. *Carbohydr. Res.* **2006**, *341*, 903–908.
- (24) Baek, J.-Y.; Choi, T. J.; Jeon, H.-B.; Kim, K.-S. A Highly Reactive and Stereoselective  $\beta$ -Mannopyranosylation System: Mannosyl 4-Pentenoate/PhSeOTf. *Angew. Chem., Int. Ed.* **2006**, *45*, 7436–7440.
- (25) Lucero, C. G.; Woerpel, K. A. Stereoselective C-Glycosylation Reactions of Pyranoses: The Conformational Preference and Reactions of the Mannosyl Cation. *J. Org. Chem.* **2006**, *71*, 2641–2647.
- (26) Smith, D. M.; Woerpel, K. A. Electrostatic Interactions in Cations and Their Importance in Biology and Chemistry. *Org. Biomol. Chem.* **2006**, *4*, 1195–1201.
- (27) Jensen, H. H.; Bols, M. Stereoelectronic Substituent Effects. *Acc. Chem. Res.* **2006**, *39*, 259–265.
- (28) Nokami, T.; Shibuya, A.; Tsuyama, H.; Suga, S.; Bowers, A. A.; Crich, D.; Yoshida, J. Electrochemical Generation of Glycosyl Triflate Pools. *J. Am. Chem. Soc.* **2007**, *129*, 10922–10928.
- (29) Park, J.; Kawatkar, S.; Kim, J.-H.; Boons, G.-J. Stereoselective Glycosylations of 2-Azido-2-deoxy-glucosides Using Intermediate Sulfonium Ions. *Org. Lett.* **2007**, *9*, 1959–1962.
- (30) Crich, D.; Sharma, I. Is Donor-Acceptor Hydrogen Bonding Necessary for 4,6-O-Benzylidene-directed  $\beta$ -Mannopyranosylation? Stereoselective Synthesis of  $\beta$ -C-Mannopyranosides and  $\alpha$ -C-Glucopyranosides. *Org. Lett.* **2008**, *10*, 4731–4734.
- (31) Manabe, S.; Ito, Y. Optimizing Glycosylation Reaction Selectivities by Protecting Group Manipulation. *Curr. Bioactive Compounds* **2008**, *4*, 258–281.
- (32) Krumer, J. R.; Salamant, W. A.; Woerpel, K. A. Continuum of Mechanisms for Nucleophilic Substitutions of Cyclic Acetals. *Org. Lett.* **2008**, *10*, 4907–4910.
- (33) Walvoot, M. T. C.; Lodder, G.; Mazurek, J.; Overkleeft, H. S.; Cde, J. D. C.; van der Marel, G. A. Equatorial Anomeric Triflates from Mannuronic Acid Esters. *J. Am. Chem. Soc.* **2009**, *131*, 12080–12081.
- (34) Zhu, X.; Schmidt, R. R. New Principles for Glycoside-Bond Formation. *Angew. Chem., Int. Ed.* **2009**, *48*, 1900–1934.
- (35) Boltje, T. J.; Buskas, T.; Boons, G.-J. Opportunities and Challenges in Synthetic Oligosaccharide and Glycoconjugate Research. *Nature Chem.* **2009**, *1*, 611–622.
- (36) Yang, M. T.; Woerpel, K. A. The Effect of Electrostatic Interactions on Conformational Equilibria of Multiply Substituted Tetrahydropyran Oxocarbenium Ions. *J. Org. Chem.* **2009**, *74*, 545–553.
- (37) Krumper, J. R.; Salamant, W. A.; Woerpel, K. A. Correlations Between Nucleophilicities and Selectivities in the Substitutions of Tetrahydropyran Acetals. *J. Org. Chem.* **2009**, *74*, 8039–8050.
- (38) Baek, J. Y.; Lee, B.-Y.; Jo, M. Gi.; Kim, K. S.  $\beta$ -Directing Effect of Electron-Withdrawing Groups at O-3, O-4, and O-6 Positions and  $\alpha$ -Directing Effect by Remote Participation of 3-O-Acyl and 6-O-Acetyl Groups of Donors in Mannopyranosylations. *J. Am. Chem. Soc.* **2009**, *131*, 17705–17713.
- (39) Stalford, S. A.; Kilner, C. A.; Leach, A. G.; Turnbull, W. B. Neighboring Group Participation vs. Addition to Oxocarbenium Ions: Studies on The Synthesis of Mycobacterial Oligosaccharides. *Org. Biomol. Chem.* **2009**, *7*, 4842–4852.
- (40) Post, C. B.; Karplus, M. Does Lysozyme Follow the Lysozyme Pathway? An Alternative Based on Dynamic, Structural, and



- Stereoelectronic Consideration. *J. Am. Chem. Soc.* **1986**, *108*, 1317–1319.
- (41) Andrews, C. W.; Fraser-Raid, B.; Bowen, J. P. An ab Initio Study (6-31G\*) of Transition States in Glycoside Hydrolysis Based on Axial and Equatorial 2-Methoxytetrahydropyrans. *J. Am. Chem. Soc.* **1991**, *113*, 8293–8298.
- (42) Woods, R. J.; Andrews, C. W.; Bowen, J. P. Molecular Mechanical Investigations of the Properties of Oxocarbenium Ions. 2. Application to Glycoside Hydrolysis. *J. Am. Chem. Soc.* **1992**, *114*, 859–864.
- (43) Andrew, C. W.; Rodebaugh, R.; Fraser-Raid, B. A Solvation-Assisted Model for Estimating Anomeric Reactivity. Predicted versus Observed Trends in Hydrolysis of n-Pentenyl Glycosides. *J. Org. Chem.* **1996**, *61*, 5280–5289.
- (44) Bérces, A.; Enright, G.; Nukada, T.; Whitfield, D. M. The Conformational Origin of the Barrier to the Formation of Neighboring Group Assistance in Glycosylation Reactions: A Dynamical Density Functional Theory Study. *J. Am. Chem. Soc.* **2001**, *123*, 5460–5464.
- (45) Bérces, A.; Whitfield, D. M.; Nukada, T. Quantitative Description of Six-Membered Ring Conformations Following the IUPAC Conformational Nomenclature. *Tetrahedron* **2001**, *57*, 477–491.
- (46) Nukada, T.; Bérces, A.; Whitfield, D. M. Can The Stereochemical Outcome of Glycosylation Reactions Be Controlled by The Conformational Preferences of The Glycosyl Donor. *Carbohydr. Res.* **2002**, *337*, 765–774.
- (47) Stubbs, J. M.; Marx, D. Glycosidic Bond Formation in Aqueous Solution: On the Oxocarbenium Intermediate. *J. Am. Chem. Soc.* **2003**, *125*, 10960–10962.
- (48) Bérces, A.; Whitfield, D. M.; Nukada, T.; do Santos, Z. I.; Obuchowska, A.; Krepinsky, J. J. Glycosylation: Is Acyl Migration to The Aglycon Avoidable in 2-Acyl Assisted Reactions. *Can. J. Chem.* **2004**, *82*, 1157–1171.
- (49) Nukada, T.; Bérces, A.; Wang, L.; Zgierski, M. Z.; Whitfield, D. M. The Two-conformer Hypothesis: 2,3,4,6-tetra-O-methylmannopyranosyl And -glucopyranosyl Oxocarbenium Ions. *Carbohydr. Res.* **2005**, *340*, 841–852.
- (50) Denekamp, C.; Sandlers, Y. Formation and Stability of Oxocarbenium Ions from Glycosides. *J. Mass Spectrom.* **2005**, *40*, 1055–1063.
- (51) Stubbs, J. M.; Marx, D. Aspects of Glycosidic Bond Formation in Aqueous Solution: Chemical Bonding and the Role of Water. *Chem.—Eur. J.* **2005**, *11*, 2651–2659.
- (52) Ionescu, A. R.; Whitfield, D. M.; Zierskia, M. Z.; Nukada, T. Investigations into The Role of Oxocarbenium Ions in Glycosylation Reactions by ab Initio Molecular Dynamics. *Carbohydr. Res.* **2006**, *341*, 2912–2920.
- (53) Whitfield, D. M.; Nukada, T. DFT Studies of The Role of C-2-O-2 Bond Rotation in Neighboring-group Glycosylation Reactions. *Carbohydr. Res.* **2007**, *342*, 1291–1304.
- (54) Biarnés, X.; Ardèvol, A.; Planas, A.; Rovira, C.; Laio, A.; Parrinello, M. The Conformational Free Energy Landscape of  $\beta$ -D-Glucopyranose. Implications for Substrate Preactivation in  $\beta$ -Glucoside Hydrolases. *J. Am. Chem. Soc.* **2007**, *129*, 10686–10693.
- (55) Whitfield, D. M. Computational Studies of the Role of Glycopyranosyl Oxocarbenium Ions in Glycobiology and Glycochemistry. *Adv. Carbohydr. Chem. Biochem.* **2009**, *62*, 83–159.
- (56) Codée, J. D. C.; van den Bos, L. J.; de Jong, A.-R.; Dinkelaar, J.; Lodder, G.; Overkleeft, H. S.; van der Marel, G. A. The Stereodirecting Effect of the Glycosyl C5-Carboxylate Ester: Stereoselective Synthesis of  $\beta$ -Mannuronic Acid Alginates. *J. Org. Chem.* **2009**, *74*, 38–47.
- (57) Pougny, J.-R.; Sinaÿ, P. Reaction d'imidates de glucoyranosyle avec l'acetonitrile. Applications synthétiques. *Tetrahedron Lett.* **1976**, *17*, 4073–4076.
- (58) Lemieux, R. U.; Ratcliffe, R. M. The Azidonitration of Tri-O-acetyl-D-galactal. *Can. J. Chem.* **1979**, *57*, 1244–1251.
- (59) Schmidt, R. R.; Rücker, E. Stereoselective Glycosidations of Uronic Acids. *Tetrahedron Lett.* **1980**, *21*, 1421–1424.
- (60) Paulsen, H. Advances in Selective Chemical Syntheses of Complex Oligosaccharides. *Angew. Chem., Int. Ed.* **1982**, *21*, 155–173.
- (61) Hashimoto, S.; Hayashi, M.; Noyori, R. Glycosylation Using Glucopyranosyl Fluorides And Silicon-based Catalysts. Solvent Dependency of the Stereoselection. *Tetrahedron Lett.* **1984**, *25*, 1379–1382.
- (62) Braccini, I.; Derouet, C.; Esnault, J.; Herve du Penhoat, C.; Mallet, J.-M.; Michon, V.; Sinaÿ, P. Conformational Analysis of Nitrilium Intermediates in Glycosylation Reactions. *Carbohydr. Res.* **1993**, *246*, 23–41.
- (63) Uchiro, H.; Mukaiyama, T. Trityl Salt Catalyzed Stereoselective Glycosylation of Alcohols with 1-Hydroxyribofuranose. *Chem. Lett.* **1996**, *1*, 79–80.
- (64) Demchenko, A.; Stauch, T.; Boons, G. J. Solvent and Other Effects on the Stereoselectivity of Thioglycoside Glycosidations. *Synlett* **1997**, 818–820.
- (65) Manabe, S.; Ito, Y.; Ogawa, T. Solvent Effect in Glycosylation Reaction on Polymer Support. *Synlett* **1998**, 628–630.
- (66) Rencurosi, A.; Lay, L.; Russo, G.; Caneva, E.; Poletti, L. Glycosylation with Trichloroacetimidates in Ionic Liquids: Influence of the Reaction Medium on the Stereochemical Outcome. *J. Org. Chem.* **2005**, *70*, 7765–7768.
- (67) Koshiba, M.; Suzuki, N.; Arihara, R.; Tsuda, T.; Nambu, H.; Nakamura, S.; Hashimoto, S. Catalytic Stereoselective Glycosidation with Glycosyl Diphenyl Phosphates: Rapid Construction of 1,2-cis- $\alpha$ -Glycosidic Linkages. *Chem. Asian J.* **2008**, *3*, 1664–1677.
- (68) Whitfield, D. M. DFT Studies of the Ionization of Alpha and Beta Glycopyranosyl Donors. *Carbohydr. Res.* **2007**, *342*, 1726–1740.
- (69) Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (70) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **1994**, *98*, 11623–11627.
- (71) 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.
- (72) Tomasi, J.; Mennucci, B.; Cammi, R. Quantum Mechanical Continuum Solvation Models. *Chem. Rev.* **2005**, *105*, 2999–3093.
- (73) Lemieux, R. U. Effects of Unshared Pairs of Electrons and Their Solvation on Conformational Equilibria. *Pure Appl. Chem.* **1971**, *25*, 527–548.
- (74) *Gauss View*, version 3.0.; Gaussian, Inc.: Pittsburgh, PA, 2003.
- (75) Christen, M.; Hünenberger, P. H.; Bakowies, D.; Baron, R.; Bürgi, R.; Geerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Kräutler, V.; Oostenbrink, C.; Peter, C.; Tryesniak, D.; van Gunsteren, W. F. The GROMOS Software for Biomolecular Simulation: GROMOS05. *J. Comput. Chem.* **2005**, *26*, 1719–1751.
- (76) 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*, 1656–1676.
- (77) Oostenbrink, C.; Soares, T. A.; van der Vegt, N. F. A.; van Gunsteren, W. F. Validation of the 53A6 GROMOS Force Field. *Eur. Biophys. J.* **2005**, *34*, 273–284.
- (78) Lins, R. D.; Hünenberger, P. H. A New GROMOS Force Field for Hexopyranose-Based Carbohydrates. *J. Comput. Chem.* **2005**, *26*, 1400–1412.
- (79) Pereira, C. S.; Kony, D.; Baron, R.; Müller, M.; van Gunsteren, W. F.; Hünenberger, P. H. Conformational and Dynamical Properties of Disaccharides in Water: A Molecular Dynamics Study. *Biophys. J.* **2006**, *90*, 4337–4344.
- (80) Pereira, C. S.; Kony, D.; Baron, R.; Müller, M.; van Gunsteren, W. F.; Hünenberger, P. H. Erratum to “Conformational and dynamical properties of disaccharides in water: A molecular dynamics study. *Biophys. J.* **2007**, *93*, 706–707.
- (81) Kräutler, V.; Müller, M.; Hünenberger, P. H. Conformation, Dynamics, Solvation and Relative Stabilities of Selected  $\beta$ -Hexopyranoses in Water: A Molecular Dynamics Study with the GROMOS 45A4 Force Field. *Carbohydr. Res.* **2007**, *342*, 2097–2124.
- (82) Hansen, H. S.; Hünenberger, P. H. Using the Local Elevation Method to Construct Optimized Umbrella Sampling Potentials: Calculation of the Relative Free Energies and Interconversion Barriers of Glucopyranose Ring Conformers in Water. *J. Comput. Chem.* **2010**, *31*, 1–23.
- (83) *The Merck Index*, 13th ed.; Merck & Co., Inc.: Whitehouse Station, NJ.
- (84) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Di Nola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (85) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of *n*-Alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.
- (86) van Gunsteren, W. F.; Berendsen, H. J. C. Computer Simulation of Molecular Dynamics: Methodology, Applications and Perspectives in Chemistry. *Angew. Chem., Int. Ed.* **1990**, *29*, 992–1023.
- (87) Sigma-Aldrich. <http://www.sigmaaldrich.com/chemistry/solvents/toluene-center.html> (accessed April 13, 2010).
- (88) Yasumi, M.; Shirai, M. The Dielectric Constant of 1,4-Dioxane. *Bull. Chem. Soc. Jpn.* **1955**, *28*, 193–196.
- (89) Rao, V. S. R.; Qasba, P. K.; Balaji, P. V.; Chandrasekaran, R. *Conformation of Carbohydrates*; Harwood Academic Publishers: Amsterdam, The Netherlands, 1998.
- (90) Romero, J. A. C.; Tabacco, S. A.; Woerpel, K. A. Stereochemical Reversal of Nucleophilic Substitution Reactions Depending upon Substituent: Reactions of Heteroatom-Substituted Six-Membered-Ring Oxocarbenium Ions through Pseudoaxial Conformers. *J. Am. Chem. Soc.* **2000**, *122*, 168–169.
- (91) Ionescu, A.; Wang, L.; Zgierski, M. Z.; Nukada, T.; Whitfield, D. M. Two Unexpected Effects Found with 2,3,4,6-Tetra-O-methyl-D-Gluc- and Mannopyranosyl Oxocarbenium Ions. In *NMR Spectroscopy and Computer Modeling of Carbohydrates*; Vliegthar, H., Woods, R. J., Eds.; *Am. Chem. Soc. Symp. Ser.*, 2006, *930*, 302–319.
- (92) Ayala, L.; Lucero, C. G.; Romero, J. A. C.; Tabacco, S. A.; Woerpel, K. A. Stereochemistry of Nucleophilic Substitution Reactions Depending upon Substituent: Evidence for Electrostatic Stabilization of Pseudoaxial Conformers of Oxocarbenium Ions by Heteroatom Substituents. *J. Am. Chem. Soc.* **2003**, *125*, 15521–15528.
- (93) Crich, D.; Vinod, A. U. Oxazolidinone Protection of *N*-Acetylglucosamine Confers High Reactivity on the 4-Hydroxy Group in Glycosylation. *Org. Lett.* **2003**, *5*, 1297–1300.
- (94) Crich, D.; Vinod, A. U. 6-*O*-Silyl-*N*-acetyl-2-amino-2-*N*,3-*O*-carbonyl-2-deoxyglucosides: Effective Glycosyl Acceptors in the Glucosamine 4-OH Series. Effect of Anomeric Stereochemistry on the Removal of the Oxazolidinone Group. *J. Org. Chem.* **2005**, *70*, 1291–1296.
- (95) Manabe, S.; Ishii, K.; Ito, Y. *N*-Benzyl-2,3-oxazolidinone as a Glycosyl Donor for Selective  $\alpha$ -Glycosylation and One-Pot Oligosaccharide Synthesis Involving 1,2-*cis*-Glycosylation. *J. Am. Chem. Soc.* **2006**, *128*, 10666–10667.
- (96) Crich, D.; Cai, F.; Yang, F. A Stable, Commercially Available Sulfonyl Chloride for the Activation of Thioglycosides in Conjunction with Silver Trifluoromethanesulfonate. *Carbohydr. Res.* **2008**, *343*, 1858–1862.
- (97) Lemieux, R. U. Newer Developments in the Conformational Analysis of Carbohydrates. *Pure Appl. Chem.* **1971**, *25*, 527–547.
- (98) Lemieux, R. U.; Hendriks, K. B.; Stick, R. V.; James, K. Halide Ion Catalyzed Glycosidation Reactions. Syntheses of  $\alpha$ -Linked Disaccharides. *J. Am. Chem. Soc.* **1975**, *97*, 4056–4062.

## Molecular Dynamics with Multiple Time Scales: How to Avoid Pitfalls

Joseph A. Morrone,<sup>†</sup> Ruhong Zhou,<sup>‡</sup> and B. J. Berne<sup>\*,†,‡</sup>

*Department of Chemistry, Columbia University, 3000 Broadway, MC 3103, New York, New York 10027, and IBM Thomas J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, New York 10598*

Received January 27, 2010

**Abstract:** Multiple time scale methodologies have gained widespread use in molecular dynamics simulations and are implemented in a variety of ways across numerous packages. However, performance of the algorithms depends upon the details of the implementation. This is particularly important in the way in which the nonbonded interactions are partitioned. In this work, we show why some previous implementations give rise to energy drifts, and how this can be corrected. We also provide a recipe for using multiple time step methods to generate stable trajectories in large scale biomolecular simulations, where long trajectories are needed.

### 1. Introduction

Molecular dynamics is a ubiquitous tool for simulating a wide variety of large scale systems, ranging from the materials to the biological sciences. Schemes that increase the efficiency of such simulations are of great interest. In standard techniques, the time step of the generated trajectory is limited by the fastest motions present in the system. However, realistic systems contain a broad spectrum of frequencies. Multiple time scale (MTS) methods partition the computation into “slow” and “fast” portions, assigning appropriate time steps to each segment. This methodology may be exploited in systems with disparate masses,<sup>1</sup> high frequency oscillators in slowly evolving baths,<sup>2</sup> and distance based schemes that partition the nonbonded interactions into short- and long-range components.<sup>3–5</sup>

The reversible reference system propagator algorithm (r-RESPA)<sup>6</sup> is one of the most powerful implementations of the multiple time scale concept. r-RESPA integrators are readily derived from factorization of the Liouville propagator.<sup>6,7</sup> It therefore provides an integration scheme that is reversible in time and evolves in a symplectic and area preserving fashion, thereby preserving these attributes of an exact solution to Hamilton’s classical equations of motion. Furthermore, a variety of different multiple time scale

partitionings may be readily recovered from this framework, including related integrators.<sup>5</sup> This algorithm has been widely implemented in simulation packages such as IMPACT,<sup>8</sup> NAMD2,<sup>9</sup> AMBER,<sup>10</sup> and DESMOND.<sup>11</sup>

The widespread availability of fast multicore computer clusters, massively parallel supercomputers, and the improvements of parallel algorithms have facilitated the simulation of longer trajectories on the order of tens of nanoseconds to microseconds for large biomolecular systems. Since the majority of papers reporting tests of the stability of multiple times scale methods were published before such advances, it is important to evaluate the validity of MTS algorithms using much longer times scales. A more recent work by Han et al.<sup>12</sup> studies disparate time scales in a simulation of a biomolecular system in a Langevin bath over several nanoseconds. Here, we focus on the stability of the integrator as measured by its energy conservation in the microcanonical simulation. This is due to the fact that coupling the system to the thermostats and barostats necessary to generate other ensembles may obscure or complicate the evaluation of the integrator’s stability.

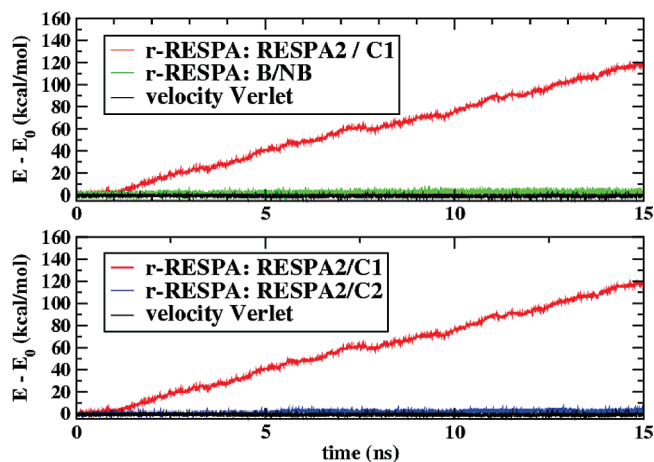
Symplectic integrators such as r-RESPA can successfully generate long stable trajectories. In addition, for such integrators, there exists a modified or shadow Hamiltonian<sup>13,14</sup> which is exactly conserved as the system is propagated, although it is only known approximately for realistic systems.<sup>15,16</sup> However, despite these desirable properties, r-RESPA and related integrators are known to suffer from

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

<sup>†</sup> Columbia University.

<sup>‡</sup> IBM Thomas J. Watson Research Center.





**Figure 1.** (Top) Deviation of the total energy from the initial energy ( $E_0 = -240\,068$  kcal/mol) plotted against trajectory length for a solvated lysozyme using the NAMD2 package. The lengths of all covalent bonds to hydrogen are constrained in all runs. The default RESPA2/C1 (red line) and BONDED/NON-BONDED (labeled B/NB, green line) schemes are plotted against a standard velocity Verlet run (black line). The bottom panel shows a NAMD2 simulation of the same system. The red and black lines correspond to the same plots as in the top panel. The blue line utilizes the RESPA2 type splitting with our fix, which has been implemented as the “long-Splitting=c2” option in NAMD2, version 2.7b2 (see section 3).

resonance instabilities,<sup>5,17–19</sup> which bound the size of the time step of the slowest motions relative to the size of the faster modes. Resonance phenomena engender the building up of energy in the system, thereby giving rise to drifts in average properties and inaccurate sampling. Schemes have been developed to alleviate this problem,<sup>18,20–27</sup> although some are only suitable for sampling and not investigations of system dynamics.

In this work, however, we do not focus on resonance phenomena in particular, but rather on optimizing the splitting of the long-range nonbonded electrostatics in order to ensure long time numerical stability. In periodic systems, nonbonded interactions may be split in two ways. We will refer to these schemes as RESPA1 (split by real-space and k-space in Ewald summation) and RESPA2 (split by intrinsic time scale; see section 2). Another consideration is how the interactions are smoothed at the boundary between the long- and short-range interactions.<sup>4,6</sup> This is facilitated by use of a switching function, the details of which may be crucial to generating a stable trajectory.

As with any numerical scheme, the performance of r-RESPA depends on the details of its implementation. In order to illustrate this point, we present in Figure 1 the deviation of the total energy from its initial energy over the course of a microcanonical simulation of lysozyme in an 8 M urea solution performed utilizing several integration schemes, as implemented in the NAMD2 simulation package. The simulations were run for up to 200 ns, but only the first 15 ns of data are shown. It can be seen that the standard velocity Verlet integrator is stable with a time step of 1 fs. In the r-RESPA schemes, the bonded interaction is evaluated with a time step of 1 fs. However, the nonbonded interaction

is handled in two different ways, each utilizing an additional time scale of 2 fs. In one version, the entire nonbonded interaction is evaluated at the larger time step. We refer to this as “RESPA:B/NB”, and the results are reported in the top panel of Figure 1. Also plotted in the top panel is the case when the nonbonded interactions are split across the two time scales according to the default implementation in NAMD2. This is denoted as “RESPA2:C1” (as we will discuss in section 2, this implementation in NAMD2 is similar to RESPA2 as presented in the literature<sup>28,29</sup> but is not exactly the same). It can be seen that, whereas RESPA:B/NB is relatively stable, a significant energy drift is present in the RESPA2:C1 result. Counterintuitively, therefore, the larger drift is seen in the case where a greater portion of the interaction is integrated at the shorter time step.

It has been shown that seemingly reasonable implementations can give rise to unexpected energy drifts. We are therefore motivated to test the stability of the various schemes of applying the r-RESPA algorithm in microcanonical simulations over long trajectories. In particular, we are interested in investigating the details of the partitioning of the nonbonded interactions. Stability depends upon the details of the interaction split, such that the incorrect choice of parameters can lead to unstable trajectories. We find that the large energy drift of RESPA2:C1 as shown in the top panel of Figure 1 is engendered by the choice of switching function that facilitates the partitioning of the nonbonded interactions (electrostatic interactions to be specific). The bottom panel of this figure shows the total energy for the trajectory for the same RESPA2 scheme, except that the switching function has been “fixed” (denoted RESPA2:C2). This modification has been recently ported into NAMD2, and its details will be explained below.

This article is organized as follows: Section 2 reviews the different ways to decompose the nonbonded interactions. In section 3, the sensitivity of the nonbonded splittings is tested for a simple water system, and what is learned here is applied to a biomolecular system in section 4. Conclusions are given in section 5.

## 2. Choosing the Force Splitting

Of crucial importance to the nature of the algorithm is the way in which the multiple time scales are defined. In a typical empirical potential, this may be done by splitting the force into a set of terms that are evaluated at different time steps. It is typical for the fastest motions to be chosen as the stretching and bending terms of the force field. The torsional terms of the potential may be included here, or treated at another level “outside” the stretching and bending interaction. The nonbonded interaction may be split into two or more parts according to their relative intrinsic time scales (fast or slow, on the basis of pair distances). In this work, we will only consider nonbonded potentials of the following form which act in a periodic simulation cell with vector of periodicity  $\mathbf{n}$ :

$$V = \sum_{\mathbf{n}} \sum_i \sum_{j \geq i} (1 - \delta_{ij}^0) \left[ \phi(r_{ij}^{\mathbf{n}}) + \frac{q_i q_j}{r_{ij}^{\mathbf{n}}} \right] \quad (1)$$

where the sum is over all periodic images and all pairs that do not correspond to the same atomic site. The function,  $\phi$ , is a short-ranged potential effectively accounting for repulsive and dispersion interactions (typically taken to be of the Lennard-Jones form), and the second term is the electrostatic interactions of fixed point charges. The electrostatic potential is long-range and may be treated via the Ewald summation technique<sup>3,30,31</sup>

$$\sum_{\mathbf{n}} \sum_i \sum_{j \geq i} (1 - \delta_{ij}^0) \frac{q_i q_j}{r_{ij}^n} = V_{\text{scr}} + V_{\text{KS}} \quad (2)$$

where  $V_{\text{scr}}$  is a short-ranged, screened potential and  $V_{\text{KS}}$  is a smooth, slowly varying potential that is most efficiently computed in reciprocal space:

$$V_{\text{scr}} = \sum_{\mathbf{n}} \sum_i \sum_{j \geq i} (1 - \delta_{ij}^0) q_i q_j \frac{\text{erfc}(\alpha r_{ij}^n)}{r_{ij}^n} \quad (3)$$

$$V_{\text{KS}} = \sum_{\mathbf{n}} \sum_i \sum_{j \geq i} (1 - \delta_{ij}^0) q_i q_j \frac{\text{erf}(\alpha r_{ij}^n)}{r_{ij}^n} \quad (4)$$

$$= \frac{1}{L^3} \sum_{\mathbf{k} \neq 0} \frac{2\pi}{k^2} e^{-k^2/4\alpha^2} |S(\mathbf{k})|^2 - V_{\text{self}} \quad (5)$$

where  $S(\mathbf{k}) = \sum_i^N q_i e^{i\mathbf{k} \cdot \mathbf{r}_i}$ . The term  $V_{\text{self}}$  subtracts out the interaction between the same sites that is implicit in the reciprocal space sum. As it is position-independent, it will not contribute to the forces and will be neglected for the rest of this discussion. The parameter  $\alpha$  determines the degree of screening and is chosen in accordance with the real space interaction cutoff  $r_{\text{cut}}$ . The reciprocal space part of the Ewald summation may be computed directly or by means of methods that utilize fast Fourier transforms (FFT) such as particle mesh Ewald (PME),<sup>32</sup> smooth particle mesh Ewald (SPME),<sup>33</sup> particle-particle particle-mesh Ewald<sup>34</sup> (P3ME), and the fast multipole method<sup>35</sup> (FMM). Such techniques have been implemented alongside r-RESPA<sup>29,36–38</sup>

Within the multiple time scale framework, the nonbonded forces may be split into two or more partitions.<sup>29,39,40</sup> In this study, we will restrict ourselves to splitting the non-bonded force into two parts. When choosing a splitting for the forces into near and far contributions, a natural choice would be to utilize the explicitly short-ranged potentials as the near force and the reciprocal space sum as the long force. This choice also has the utility that the more computationally expensive reciprocal space part is computed less frequently. Following the nomenclature of ref 29, this choice is referred to as “RESPA1.”

$$V_{\text{near}}^{(1)} = \sum_{\mathbf{n}} \sum_i \sum_{j \geq i} (1 - \delta_{ij}^0) [\phi(r_{ij}^n) + q_i q_j \frac{\text{erfc}(\alpha r_{ij}^n)}{r_{ij}^n}] \times (1 - \Theta(r_{ij}^n - r_{\text{cut}}))$$

$$V_{\text{far}}^{(1)} = V_{\text{KS}} \quad (6)$$

The factor  $(1 - \Theta(r_{ij}^n - r_{\text{cut}}))$  cuts off the interaction in real space, where  $\Theta(r)$  is the Heavyside function. Smoother

functions may be utilized to facilitate improved energy conservation.

As noted in previous work,<sup>28,29</sup> this is not the optimal split, as some fast components are screened out of the potential in eq 3 and are therefore present in the reciprocal space term (eq 5). One may subtract this portion from the “far” potential and add it into the “near” potential, yielding a split that we will refer to (as in ref 29) as “RESPA2.”

$$V_{\text{near}}^{(2)} = V_{\text{near}}^{(1)} + \sum_{\mathbf{n}} \sum_i \sum_{j \geq i} (1 - \delta_{ij}^0) (1 - \Theta(r_{ij}^n - r_{\text{cut}})) \times \frac{q_i q_j \frac{\text{erf}(\alpha r_{ij}^n)}{r_{ij}^n}}{r_{ij}^n} \quad (7)$$

$$= \sum_{\mathbf{n}} \sum_i \sum_{j \geq i} (1 - \delta_{ij}^0) (1 - \Theta(r_{ij}^n - r_{\text{cut}})) \times \left[ \phi(r_{ij}^n) + \frac{q_i q_j}{r_{ij}^n} \right]$$

$$V_{\text{far}}^{(2)} = V_{\text{KS}} - \sum_{\mathbf{n}} \sum_i \sum_{j \geq i} (1 - \delta_{ij}^0) (1 - \Theta(r_{ij}^n - r_{\text{cut}})) \times \frac{q_i q_j \frac{\text{erf}(\alpha r_{ij}^n)}{r_{ij}^n}}{r_{ij}^n} \quad (8)$$

The forces may be obtained by taking the negative gradient of the associated piece of the potential and then splitting it according to the r-RESPA algorithm.<sup>6</sup> In general, a cutoff different than the overall real space interaction cutoff,  $r_{\text{cut}}^{\text{res}}$ , may be employed for the division between the “near” and “far” forces. The forces may be decomposed as follows:<sup>4,6</sup>

$$f_{\text{inner}} = f_{\text{near}} S(r; r_{\text{cut}}^{\text{res}}, \lambda)$$

$$f_{\text{outer}} = f_{\text{far}} + f_{\text{near}} (1 - S(r; r_{\text{cut}}^{\text{res}}, \lambda)) \quad (9)$$

where  $S$  is a switching function that softens the transition from fast to slow forces that occurs at a given cutoff. The switching function is made to act over a healing length,  $\lambda$ . Too harsh a transition can lead to errors in the region of the cutoff, thereby introducing instabilities in the propagation that accumulate over the integration time. In some implementations,<sup>9,22</sup> the entirety of short ranged potential  $\phi$  is placed in the inner loop and only the electrostatic interactions are split. It is also possible to instead apply a switching function directly to the potential.<sup>22,41</sup> If  $S$  is applied to the potential, then the corresponding switching function  $\tilde{S}$  will act upon the force, thereby replacing  $S$  in eq 9 with the following expression:

$$\tilde{S}(r) = S(r) + \frac{V(r)}{V'(r)} S'(r) \quad (10)$$

where the prime indicates a derivative with respect to distance  $r$ . Note that, since the forces and not the potentials are utilized to generate the trajectories, it is the smoothness of the switching function that acts on the force which will impact the integration stability. In principle, equivalent switching functions, either on the potential or on the force, will generate the same smooth trajectories.



The switching function,  $S$ , used here is given in the following piecewise form:

$$S(r; r_{\text{cut}}^{\text{res}}, \lambda) = \begin{cases} 1 & r \leq r_{\text{cut}}^{\text{res}} - \lambda \\ g(r; r_{\text{cut}}^{\text{res}}, \lambda) & r_{\text{cut}}^{\text{res}} - \lambda < r < r_{\text{cut}}^{\text{res}} \\ 0 & r \geq r_{\text{cut}}^{\text{res}} \end{cases} \quad (11)$$

where the function,  $g$ , is chosen so as to smoothly transition from 1 to 0. In this work, we will consider three forms of this function: a cubic spline,  $g_3$ , that was utilized on the force in the original formulation of r-RESPA,<sup>4,6</sup> a quintic spline,  $g_5$ , that was utilized to cut off the electrostatic potential in a different context,<sup>42</sup> and a different form of cubic spline (denoted as the C1 spline),  $g_{\text{C1}}$ , which is the default choice in NAMD.<sup>9</sup>

$$g_3(r; r_{\text{cut}}^{\text{res}}, \lambda) = 1 + u^2(2u - 3) \quad (12)$$

$$g_5(r; r_{\text{cut}}^{\text{res}}, \lambda) = 1 + u^3(15u - 6u^2 - 10) \quad (13)$$

$$g_{\text{C1}}(r; r_{\text{cut}}^{\text{res}}, \lambda) = 1 + \frac{u}{2}(u^2 - 3) \quad (14)$$

where:

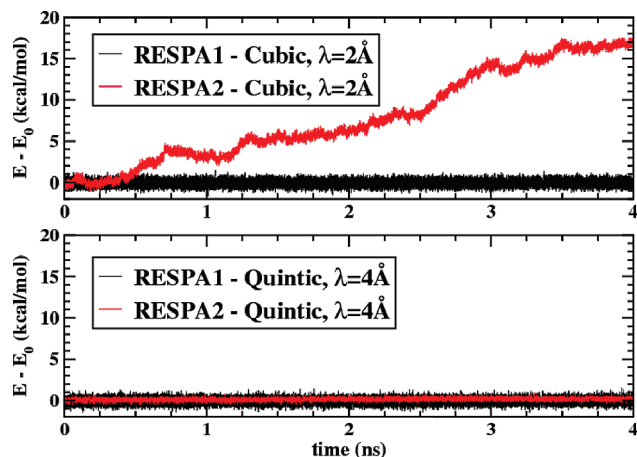
$$u = \frac{1}{\lambda}(r - r_{\text{cut}}^{\text{res}} + \lambda) \quad (15)$$

The quintic spline has the benefit of being a smoother function than the other choices, although it is also marginally more computationally expensive. Of course, we make no pretense of making the optimal choice, and other functions may be appropriate. The sensitivity of the algorithm performance to the nonbonded interaction splitting will be considered in the next section.

### 3. Testing the Splitting

In order to test the performance of the schemes delineated above, we perform a series of simulations and monitor the total energy conservation via the deviation of the energy from its initial energy,  $E - E_0$ , as a function of time, where  $E$  is the total energy and  $E_0$  is the initial energy. This plot monitors the drift and fluctuations of the conserved quantity.

We carried out simulations utilizing the PINY\_MD package.<sup>43</sup> The PINY\_MD package contains a multifaceted r-RESPA implementation to which we have added the RESPA2 splitting and the quintic switching function. The particle mesh Ewald method is utilized for computing the long-range interactions.<sup>33</sup> An overall real space cutoff of 10 Å and an  $r_{\text{cut}}^{\text{res}}$  of 8 Å were utilized in all of our PINY\_MD studies in this and the succeeding section. Using a relatively small cutoff for  $r_{\text{cut}}^{\text{res}}$  is presently computationally advantageous due to the fact that it shifts a greater burden onto the less frequently evaluated “far” interactions. It has been shown by Han et al.<sup>12</sup> that the cutoffs may be increased to yield larger differences between the inner and outer time steps such that the algorithm efficiency may be optimized according to the features of the simulation package and available hardware. The switching function is applied directly to the force as in eq 9.



**Figure 2.** The deviation,  $E - E_0$  (where  $E_0 = -7259$  kcal/mol), of the conserved energy for TIP3P water when the RESPA1 (black lines) and RESPA2 (red lines) schemes are utilized. The top panel depicts the results if a cubic switching function with a healing length of 2 Å is employed, whereas in the bottom panel a smoother choice for the switch (a quintic function with a healing length of 4 Å) is made.

Due to the large difference between fast OH stretches and the slower librations and translational motions, liquid water is a natural system on which to test the multiple time scale approach. However, the high frequency of the OH stretch induces resonance instabilities (see section 1) at rather small outer time steps.<sup>22</sup> The resonance barrier may be simply postponed by constraining the lengths of all the covalent bonds to hydrogens, and this is the approach that we follow here. Therefore, we utilize a rigid model of water. Since all covalent bond lengths are being constrained, the forces are only split between near and far nonbonded contributions.

We simulate a system of 905 water molecules in a periodic cubic cell with a side 30 Å in length, and the TIP3P model<sup>44</sup> is utilized to describe the interactions. The internal geometry of each molecule is constrained.<sup>45</sup> The near forces are updated every 1 fs, whereas the far forces are updated every 5 fs. The Ewald screening parameter utilized is  $\alpha = 0.4$  Å. In this regime, previous studies have shown that the multiple time scale splitting is stable for rigid water models.<sup>22,29</sup> Simulations of 4 ns in length were carried out in order to test the sensitivity of RESPA1 and RESPA2 to the choice of switching function and healing length.

This model will serve as our test of the sensitivity to choice of switching function parameters and functional forms. Healing lengths of 2 and 4 Å, as well as the cubic and quintic forms, were studied. In Figure 2,  $E - E_0$  is plotted versus time for selected choices of the switching function using both the RESPA1 and RESPA2 schemes. Additional stability data for all runs are given in the Appendix. It can be seen that the energy conservation dramatically improves as the switching function is made “smoother” for RESPA2, although RESPA1 is relatively insensitive to this change. The switching function can be made smoother by increasing its order and by increasing the distance over which the function acts ( $\lambda$ ). As is evident from the data in the Appendix, the careful choice of both of these aspects is necessary to optimize performance. Sensitivity of the RESPA2 scheme to the

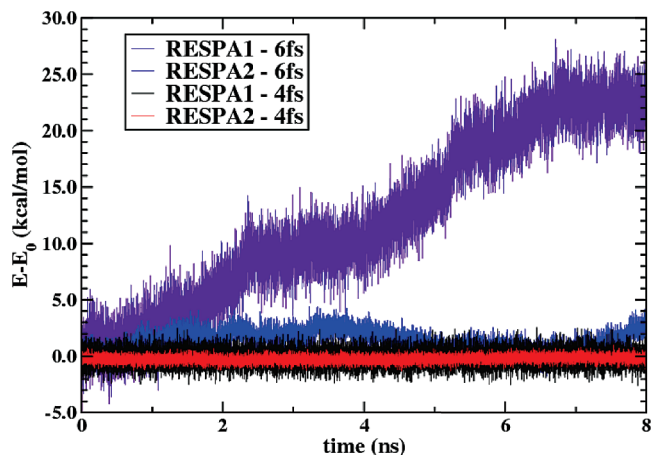
choice of healing length has been previously noted in ref 24. The RESPA1 scheme is relatively insensitive to this choice due to the fact that the interactions are already damped in the cutoff region because the screened and not the bare Coulomb potential is utilized for the “near interaction” (see eq 6). Furthermore, one may note that the drift observed in the RESPA2 scheme for a poor choice of switching function can go unnoticed over the first several hundred picoseconds of the simulation. This observation underlines the importance of monitoring longer trajectories in order to assess the performance of any integration scheme.

We now return to the original question posed by Figure 1. As noted in the Introduction, the nonbonded portion is split according to RESPA2 within NAMD2.<sup>9</sup> By default, the C1 switching function (eq 14), with a hard wired healing length of  $\lambda = r_{\text{cut}}$ , is applied directly to the *potential*. Numerical stability therefore depends upon the smoothness of  $\tilde{S}(r)$  as defined in eq 10 (see section 2). In the bottom panel of Figure 1, we plot the results shown in the top panel of this figure against what occurs if the default C1 switching function is simply replaced with our smoother quintic switching function for the RESPA2 scheme.<sup>46</sup> One can see that this single modification largely alleviates the drift in the conserved energy and provides a fix, albeit not an optimal one.

#### 4. Appropriate Settings for Biomolecular Simulations

In current research, the primary utility of multiple time scale methods is to increase the computational efficiency of biomolecular simulations. To this end, we make a careful study of the performance of r-RESPA for such systems. As a test case, we simulate a lysozyme surrounded by 6328 TIP3P water molecules in a periodic cubic box with a side of length 61.5 Å using the CHARMM22<sup>47</sup> force field. The Ewald screening parameter is set to 0.37 Å. We chose a time step of 1 fs for the near nonbonded interactions and torsional terms, and we integrate the stretching and bending terms of the protein with a time step of 0.5 fs. The outer time step that splits the nonbonded interactions is varied. We utilize a quintic switching function with a healing length of 4 Å in all runs. All water molecules are taken to be rigid, and all bonds to hydrogen (except hydrogen bonds) within the protein are also constrained, so as to delay resonance instabilities (see sections 1 and 3). We equilibrate the system using the RESPA2-4FS protocol for over 4 ns.

Several runs are presented in Figure 3. It is shown that the splitting is strictly stable up to 4 fs, whereas larger steps exhibit some degree of drifting due to the onset of resonance effects discussed in section 1. When the switching function is properly set, it can be seen that the RESPA2 scheme outperforms RESPA1. This is indicated by the smaller energy fluctuations in the RESPA2 runs. For an outer time step of 6 fs, the RESPA2 result also possesses a significantly smaller drift than the corresponding RESPA1 run and, in fact, appears to be fairly stable, as shown in Figure 3. The small drifts become more apparent, however, as the simulation progresses beyond 8 ns (results not shown).



**Figure 3.** The deviation,  $E - E_0$ , of the energy from the initial energy is plotted for selected runs of lysozyme utilizing RESPA1 with an outer time step of 4 fs (black line) and 6 fs (purple line) and RESPA1 with an outer time step of 4 fs (red line) and 6 fs (blue line). For this system,  $E_0 = -53\,377$  kcal/mol.

**Table 1.** Comparison of the Average Total Energy, Potential Energy, and Temperature for Selected Simulations of the Solvated Lysozyme System<sup>a</sup>

system	$E$ (kcal/mol)	$T$ (K)	$V$ (kcal/mol)
NO-SPLIT-1FS	-53374 (0.17)	299.45 (1.7)	-66138 (73)
RESPA1-4FS	-53373 (0.73)	299.24 (1.7)	-66127 (72)
RESPA2-4FS	-53374 (0.30)	299.64 (1.7)	-66147 (73)

<sup>a</sup> The standard deviation of each quantity is given in parentheses.

Furthermore, it is shown in Table 1 that the r-RESPA runs yield equivalent averages to runs where the nonbonded interactions are not split and evaluated every femtosecond. It may be possible to utilize larger outer time steps by either increasing the real space cutoff<sup>12</sup> or by splitting the nonbonded interaction into more than two portions,<sup>29</sup> where different time scales may be used to characterize near, intermediate, and long-range nonbonded interactions.

#### 5. Conclusion

Multiple time scale molecular dynamics techniques can be an important tool for the creation of optimized molecular dynamics integrators. Although splitting the nonbonded interactions according to their intrinsic space or time scales may be readily accomplished, the proper division of the intermediate and long-range interactions so as to ensure stability and energy conservation can be full of pitfalls. In this work, we have performed a detailed study of the accuracy of nonbonded splitting schemes, in particular, schemes where the electrostatic interactions are split into real and reciprocal space (RESPA1) parts or distance based (RESPA2) contributions. It is found that, while RESPA2 outperforms RESPA1, it has a greater dependence upon the details of the function that switches between the two contributions. This dependence can lead to rather significant drifts in the total energy over the course of long simulations, as is found in some previous implementations. To this end, we have provided some guidance for nonbonded splitting

**Table 2.** Simulation Details as well as Energy Conservation Measures  $E_{\text{conv}}$  and  $R$  for the Simulations of Water (section 3) and the Solvated Lysozyme (section 4)<sup>a</sup>

system	split	$\Delta t_{\text{outer}}$		length		$\log(E_{\text{conv}})$	$R (\times 10^{-2})$
		(fs)	$sw_{\text{ord}}$	$\lambda(\text{\AA})$	(ns)		
water	RESPA1	5	3	2.0	4.0	-4.42	1.37
water	RESPA1	5	5	4.0	4.0	-4.42	1.37
water	RESPA2	5	3	2.0	4.0	-2.96	21.8
water	RESPA2	5	5	2.0	4.0	-4.38	1.48
water	RESPA2	5	3	4.0	4.0	-3.75	5.63
water	RESPA2	5	5	4.0	4.0	-4.65	0.628
lysozyme	NO-SPLIT	1			2.0	-5.58	0.231
lysozyme	RESPA1	4	5	4.0	8.0	-4.95	1.02
lysozyme	RESPA1	6	5	4.0	8.0	-3.64	10.4
lysozyme	RESPA2	4	5	4.0	8.0	-5.23	0.408
lysozyme	RESPA2	6	5	4.0	8.0	-4.59	1.35

<sup>a</sup> Entries are categorized according to the size of the outer loop time step, the type of non-bonded splitting employed (RESPA1, RESPA2, or NO-SPLIT), the details of the switching function (order ( $sw_{\text{ord}}$ ) and healing length ( $\lambda$ )), and trajectory length.

schemes and implemented these in selected simulation packages (NAMD2 and PINY\_MD). Even though these implementations are applied to particular simulation packages, we believe that these findings are of broader applicability to multiple time scale methods and will be particularly useful for modern biomolecular simulations where long trajectories on the order of tens of nanoseconds to microseconds are needed.

**Acknowledgment.** This research was supported from a grant to B.J.B from the National Science Foundation via grant (NSF-CHE-0910943). R.Z. acknowledge support from IBM Blue Gene Science Program. We acknowledge Mark Tuckerman for his assistance with regards to our modifications of PINY\_MD. We also acknowledge Jingyuan Li and Thomas Markland for useful discussions.

## Appendix: Quantifying the Integrator Stability

We utilize two standard measures in order to assess the stability of the run:<sup>48-50</sup>

$$E_{\text{conv}} = \frac{|E - E_0|}{|E_0|} \quad (16)$$

$$R = \frac{\langle (E - \langle E \rangle)^2 \rangle^{1/2}}{\langle (T - \langle T \rangle)^2 \rangle^{1/2}} \quad (17)$$

where  $E$  is the total energy and  $T$  is the kinetic energy. In general,  $E_{\text{conv}}$  is more sensitive to the drift in energy, whereas  $R$  is more directly related to its fluctuation. The details of all the simulations performed with PINY\_MD are given alongside these measures of stability in Table 2. It can be seen that RESPA2 does not outperform RESPA1 until both a suitable smooth switching function form and appropriate healing length are chosen.

## References

- (1) Tuckerman, M.; Berne, B.; Rossi, A. *J. Chem. Phys.* **1991**, *94*, 1465.
- (2) Tuckerman, M.; Martyna, G.; Berne, B. *J. Chem. Phys.* **1990**, *93*, 1287.
- (3) Allen, M.; Tildesley, D. *Computer Simulation of Liquids*; Oxford University Press: Oxford, 1987.
- (4) Tuckerman, M.; Berne, B.; Martyna, G. *J. Chem. Phys.* **1991**, *94*, 6811.
- (5) Grubmuller, H.; Heller, H.; Windemuth, A.; Schulten, K. *Mol. Simul.* **1991**, *6*, 121.
- (6) Tuckerman, M.; Berne, B.; Martyna, G. *J. Chem. Phys.* **1992**, *97*, 1990.
- (7) Sexton, J.; Weingarten, D. *Nucl. Phys. B* **1992**, *380*, 665.
- (8) Banks, J.; Beard, H.; Cao, Y.; Cho, A.; Damm, W.; Farid, R.; Felts, A.; Halgren, T.; Mainz, D.; Maple, J.; Murphy, R.; Philipp, D.; Repasky, M.; Zhang, L.; Berne, B.; Friesner, R.; Gallicchio, E.; Levy, R. *J. Comput. Chem.* **2005**, *26*, 1752.
- (9) Phillips, J.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *2*, 1781.
- (10) Case, D.; Darden, T.; Cheatham, I.; Simmerling, C.; Wang, J.; Duke, R.; Luo, R.; Crowley, M.; Walker, R.; Zhang, W.; Merz, K.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossvary, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D.; Seetin, M.; Sagui, C.; Babin, V.; Kollman, P. *AMBER 10*; University of California: San Francisco, 2008.
- (11) Bowers, K.; Chow, E.; Xu, H.; Dror, R.; Eastwood, M.; Kolossvary, B.; Moraes, M.; Sacerdoti, F.; Salmon, J.; Shan, Y.; Shaw, D. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. Proceedings of the ACM/IEEE Conference on Supercomputing (SC06), Tampa, FL, 2006.
- (12) Han, G.; Deng, Y.; Glimm, J.; Martyna, G. *Comput. Phys. Commun.* **2007**, *176*, 271.
- (13) Skeel, R.; Zhang, G.; Schlick, T. *Siam J. Sci. Comput.* **1997**, *18*, 203.
- (14) Gans, J.; Shalloway, D. *Phys. Rev. E* **2000**, *61*, 4587.
- (15) Skeel, R.; Hardy, D. *Siam J. Sci. Comput.* **2001**, *23*, 1172.
- (16) Engle, R.; Skeel, R.; Drees, M. *J. Comput. Phys.* **2005**, *206*, 432.
- (17) Biesadecki, J.; Skeel, R. *J. Comput. Phys.* **1993**, *109*, 318.
- (18) Schlick, T.; Mandziuk, M.; Skeel, R.; Srinivas, K. *J. Comput. Phys.* **1998**, *140*, 1.
- (19) Ma, Q.; Izaguirre, J.; Skeel, R. *Siam J. Sci. Comput.* **2003**, *24*, 1951.
- (20) Barth, E.; Schlick, T. *J. Chem. Phys.* **1998**, *109*, 1617.
- (21) Sandu, A.; Schlick, T. *J. Comput. Phys.* **1999**, *151*, 74.
- (22) Izaguirre, J.; Reich, S.; Skeel, R. *J. Chem. Phys.* **1999**, *110*, 9853.
- (23) Izaguirre, J.; Catarello, D.; Wozniak, J.; Skeel, R. *J. Chem. Phys.* **2001**, *114*, 2090.
- (24) Qian, X.; Schlick, T. *J. Chem. Phys.* **2002**, *116*, 5971.
- (25) Ma, Q.; Izaguirre, J. *Multiscale Model. Simul.* **2003**, *2*, 1.
- (26) Minary, P.; Tuckerman, M.; Martyna, G. *Phys. Rev. Lett.* **2004**, *93*, 150201.
- (27) Sweet, C.; Petrone, P.; Pande, V.; Izaguirre, J. *J. Chem. Phys.* **2008**, *128*, 145101.

- (28) Stuart, S.; Zhou, R.; Berne, B. *J. Chem. Phys.* **1996**, *105*, 1426.
- (29) Zhou, R.; Harder, E.; Xu, H.; Berne, B. *J. Chem. Phys.* **2001**, *115*, 2348.
- (30) Ewald, P. *Ann. Phys.* **1921**, *64*, 253.
- (31) Frenkel, D.; Smit, B. *Understanding Molecular Simulation*, 2nd ed.; Academic Press: London, 2002.
- (32) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089.
- (33) Essmann, U.; Perera, L.; Berkowitz, M.; Darden, T.; Lee, H.; Pedersen, L. *J. Chem. Phys.* **1995**, *103*, 8577.
- (34) Luty, B.; Tironi, I.; van Gunsteren, W. *J. Chem. Phys.* **1995**, *103*, 3014.
- (35) Greengard, L.; Rokhlin, V. *J. Comput. Phys.* **1985**, *60*, 187.
- (36) Zhou, R.; Berne, B. *J. Chem. Phys.* **1995**, *103*, 9444.
- (37) Figueirido, F.; Levy, R.; Zhou, R.; Berne, B. *J. Chem. Phys.* **1997**, *106*, 9835.
- (38) Procacci, P.; Darden, T.; Marchi, M. *J. Phys. Chem.* **1996**, *100*, 10464.
- (39) Procacci, P.; Marchi, M. *J. Chem. Phys.* **1996**, *104*, 3003.
- (40) Procacci, P.; Marchi, M.; Martyna, G. *J. Chem. Phys.* **1998**, *108*, 8799.
- (41) Procacci, P.; Berne, B. *J. Chem. Phys.* **1994**, *101*, 2421.
- (42) Lau, K.; Alper, H.; Thacher, T.; Stouch, T. *J. Phys. Chem.* **1994**, *98*, 8785.
- (43) Tuckerman, M.; Yarne, D.; Samuelson, S.; Hughes, A.; Martyna, G. *Comput. Phys. Commun.* **2000**, *128*, 333.
- (44) Jorgensen, W.; Chandrasekhar, J.; Madura, J.; Impey, R.; Klein, M. *J. Chem. Phys.* **1983**, *79*, 926.
- (45) Rychaert, J.; Ciccotti, G.; Berendsen, H. *J. Comput. Phys.* **1977**, *23*, 327.
- (46) Our implementation of the quintic spline function has been ported into the official NAMD2 package, version 2.7b (<http://www.ks.uiuc.edu/Research/namd/>, accessed May 2010) and may be activated by setting "longSplitting=c2".
- (47) MacKerell, A.; Bashford, D.; Bellott, M.; Dunbrack, R.; Evanseck, J.; Field, M.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D.; Prodhom, B.; Reiher, W.; Roux, B.; Schlenkrich, M.; Smith, J.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Chem. Phys. B* **1998**, *102*, 3586.
- (48) van Gunsteren, W.; Berendsen, H. *Mol. Phys.* **1977**, *34*, 1311.
- (49) Watanabe, M.; Karplus, M. *J. Chem. Phys.* **1993**, *99*, 8063.
- (50) Humphreys, D.; Friesner, R.; Berne, B. *J. Phys. Chem.* **1994**, *98*, 6885.

CT100054K



## Milestoning without a Reaction Coordinate

Peter Májek<sup>\*,†</sup> and Ron Elber<sup>\*,‡</sup>

*Department of Computer Science, Upson Hall 4130, Cornell University, Ithaca, New York 14853-7501, and Department of Chemistry and Biochemistry, Institute for Computational Engineering and Sciences, 1 University Station, ICES, C0200, The University of Texas, Austin, Texas 78712*

Received February 26, 2010

**Abstract:** Milestoning is a method for calculating kinetics and thermodynamics of long time processes typically not accessible for straightforward Molecular Dynamics (MD) simulation. In the Milestoning approach, the system of interest is partitioned into cells by dividing hypersurfaces (Milestones) and transitions are computed between nearby hypersurfaces. Kinetics and thermodynamics are derived from the statistics of these transitions. The original Milestoning work concentrated on systems in which a one-dimensional reaction coordinate or an order parameter could be identified. In many biomolecular processes, the reaction proceeds via multiple channels or following more than a single-order parameter. A description based on a one-dimensional reaction coordinate may be insufficient. In the present paper, we introduce a variation that overcomes this limitation. Following the ideas of Vanden-Eijnden and Venturoli on Voronoi cells that avoid the use of an order parameter (*J. Chem. Phys.* 2009, 130, 194101), we describe another way to “Milestone” systems without a reaction coordinate. We examine the assumptions of the Milestoning calculations of mean first passage times (MFPT) and describe strategies to weaken these assumptions. The method described in this paper, Directional Milestoning, arranges hypersurfaces in higher dimensions that “tag” trajectories such that efficient calculations can be done and at the same time the assumptions required for exact calculations of MFPTs are satisfied approximately. In the original Milestoning papers, trajectories are initiated from an equilibrium set of conformations. Here a more accurate distribution, that mimics the first hitting point distribution, is used. We demonstrate the usage of Directional Milestoning in conformational transitions of alanine dipeptide (in vacuum and in aqueous solution) and compare the correctness, efficiency, and statistical stability of the method with exact MD and with a related method.

### 1. Introduction

Milestoning is a method to calculate kinetics and thermodynamics of molecular systems that evolve on long time scales typically not accessible for straightforward molecular dynamics (MD) simulation.<sup>1–8</sup>

Straightforward molecular dynamics can be used to compute rate of reactions. In these applications coordinates and velocities are initiated in the reactant state and the

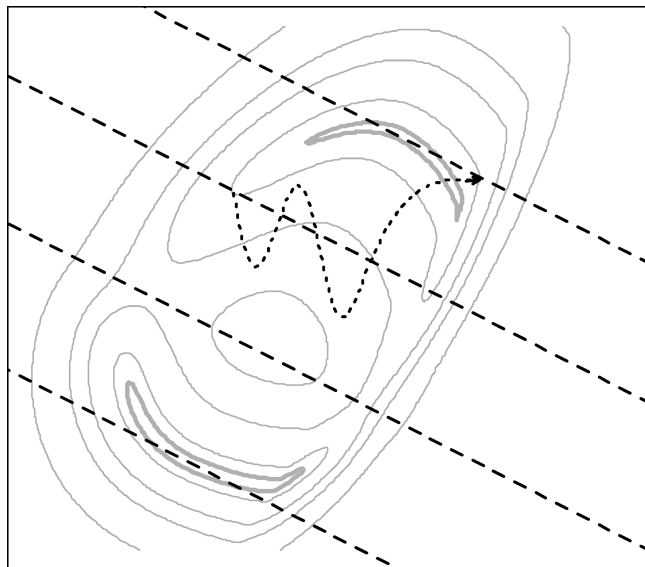
equations of motion are integrated until the product state is reached. While considerably promising, there are caveats: (i) the numerical integration of a typical biomolecular process is computationally demanding and may not be feasible; (ii) actual realizations of reactive trajectories are noisy, making their analysis difficult and may require significant filtering to recover useful signals.

In Milestoning, the conformational space between the reactant and the product is partitioned by a set of dividing hypersurfaces called Milestones (Figure 1). An ensemble of initial conditions is prepared at each Milestone and trajectories are simulated from each initial point until another nearby Milestone is reached. These trajectories are signifi-

\* To whom correspondence should be addressed. E-mail: pmajek@cs.cornell.edu (P.M.); ron@ices.utexas.edu (R.E.).

<sup>†</sup> Cornell University.

<sup>‡</sup> The University of Texas.



**Figure 1.** A schematic arrangement of Milestones (dashed lines) in a two-well potential. Also shown is a trajectory (dotted line) starting on a second Milestone and terminating on the first one.

cantly shorter and trivially parallelized compared to a reactive trajectory of the overall process. The efficiency of the algorithm is discussed in.<sup>1</sup>

In the original milestoneing papers,<sup>1,3</sup> a theory that relates the statistical properties of the short trajectories initiated on each Milestone and the overall rate was developed. In the present work we consider a variant of the Markovian limit of Milestoneing,<sup>1,2</sup> a method that uses only the first moments of local first passage time (LFPT) distributions. The advantage of the Markovian limit of Milestoneing is that it is easier to implement and is statistically more stable. As we will show in section 2.1, it calculates the overall mean first passage times (MFPT) accurately, given that certain assumptions are met. Milestoneing in its complete settings (non-Markovian) provides a useful alternative if more detailed understanding of the reaction process is desired, for example if the reaction is nonexponential in time.

Vanden-Eijnden et al.,<sup>4</sup> considered reaction dynamics with overdamped Langevin. It was shown that if Milestones are chosen as isocommittor surfaces, i.e. surfaces for which the probability of reaching the product state before the reactant is constant, then Milestoneing calculation of the MFPT using Brownian dynamics is exact. However, determination of exact isocommittor surfaces can be very difficult in practice.

Other limits in which Milestoneing is expected to be accurate are available for systems near equilibrium. As outlined in the original Milestoneing papers,<sup>1,3</sup> even when other surfaces are used (surfaces that are not isocommittors) Milestoneing can still work well. If successive crossing events of Milestones are sufficiently separated in time to “lose” velocity memory Milestoneing was illustrated to provide accurate results. This assumption is achieved in practice by placing Milestones sufficiently far from each other such that the average termination time of trajectories is at least a few hundred femtoseconds.<sup>1</sup>

In section 2, we propose a variant of Milestoneing in the Markovian limit, which we call Directional Milestoneing

(DiM), the dividing hypersurfaces are redefined in more than one dimension to capture features of the reaction (e.g., multiple reaction channels or multiple collective variables) that at the same time maintain the concept of Milestone separation, for example, trajectories initiated on any Milestone have time to “lose memory” before terminating on other Milestones.

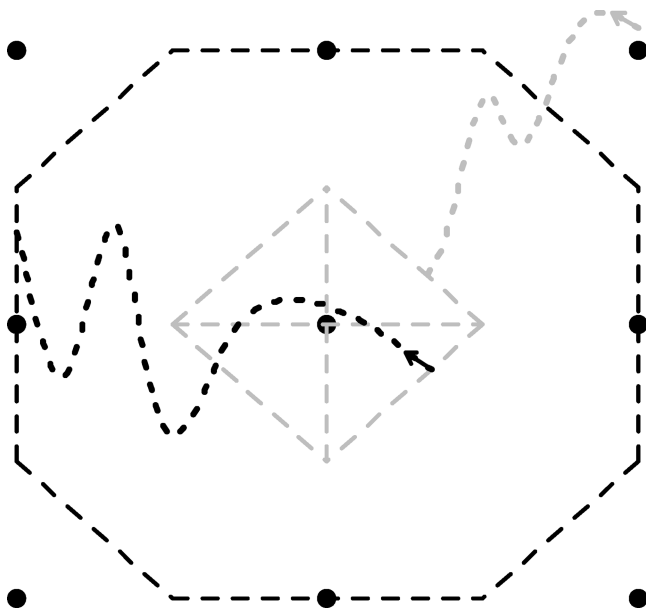
The original Milestoneing approach approximates the initial ensemble on each hypersurface by equilibrium distribution. To be exact the initial distribution at a Milestone must be the first hitting point distribution (FHPD). A first hitting point is a phase space point on the Milestone crossed for the first time by a trajectory arriving from a nearby hypersurface. The distribution of these phase space points is complex and a closed form of it is known only for overdamped Langevin dynamics in low dimensions.<sup>4</sup>

In recent work,<sup>7</sup> Vanden-Eijnden and Venturoli proposed a modification of Milestoneing that avoids a generation of initial ensembles on each of the dividing surfaces. As we discuss later their approach is more accurate compared to the original Milestoneing for the generation of the FHPD. Memory loss, however, is harder to control in the new approach. To improve the accuracy of the original Milestoneing approach, while retaining some of its advantages we propose in section 2.4 another way to approximate FHPD, which is better than the original Milestoneing.

In section 3, we illustrate the Directional Milestoneing (DiM) for the calculation of MFPT of a conformational transition of alanine dipeptide, both in vacuum and in water. We compare Directional Milestoneing with exact Molecular Dynamics and with the related method Markovian Milestoneing with Voronoi Tessellation (MMVT).<sup>7</sup> We illustrate that as the complexity of the underlying energy surface increases, DiM becomes more effective. Discussions and conclusions are in presented section 4.

## 2. Directional Milestoneing: Theory

**2.1. Definition of Milestones in Higher Dimensions.** We discuss below an extension of Milestoneing that avoids the use of a reaction coordinate. Instead of placing hypersurfaces orthogonal to a one-dimensional curve as introduced in the original papers,<sup>1,3</sup> we define the interfaces (Milestones) based on a set of coordinates (images) that sample the conformational space of the biophysical process under consideration. (Two of the images define the reactant and the product state.) These images may be obtained from long time simulations, high temperature trajectories, replica exchange simulations, etc., as discussed later in examples in the article. Having  $N$  images  $X_1, \dots, X_N$  placed in the conformational space, we intuitively want to arrange Milestones as interfaces between the images, which is the approach taken in the Voronoi Tessellation of Markovian Milestoneing.<sup>7</sup> However, we aim to place the Milestones in conformational space in such a way that a trajectory initiated on any Milestone has time and space to “lose memory” of its starting point before terminating at a different Milestone. A formal definition of “losing memory” will be given in the following section. For each pair of images  $X_i$  and  $X_j$ , we define the Milestone  $M_{i \rightarrow j}$



**Figure 2.** Example of Milestones according to definition 1. Conformational images are represented as black dots; Milestones related to the central image are displayed as dashed lines. A trajectory coming to the central region (gray, dotted) terminates on one of the gray Milestones (depending on the previously assigned region). A trajectory reinitiated on any of the gray (incoming) Milestones leaves the region through one of the black (outgoing) Milestones.

as a set of conformational points on which a trajectory enters the region of image  $X_j$  from the region of image  $X_i$ . Formally, the above intuitive requirements on Milestone placement can be accomplished in several different ways. We define a Milestone  $M_{i \rightarrow j}$  as

$$M_{i \rightarrow j} \equiv \{X | d(X, X_i)^2 = d(X, X_j)^2 + \Delta_i^2 \text{ and } \forall k d(X, X_j) \leq d(X, X_k)\} \quad (1)$$

where  $d(X, Y)$  is a distance function of images  $X$  and  $Y$  and  $\Delta_i = \min_{j \neq i} d(X_i, X_j)$ . The arrangement (eq 1) has a few important properties discussed in detail in section 2.3. We name some of the properties here, referring the formal proofs to section 2.3: A Milestone  $M_{i \rightarrow j}$  is located in the region between the images  $X_i$  and  $X_j$  and is always closer to the image  $X_j$ . The Milestone  $M_{i \rightarrow j}$  does not intersect any of  $M_{i \rightarrow l}$  Milestones (for  $l \neq j$ ), and there is a finite separation in conformational space between the Milestones  $M_{i \rightarrow j}$  and  $M_{l \rightarrow i}$ . See Figure 2 for an example of the proposed arrangement. As shown in the figure, the outgoing (black) Milestones bound the region of the central image and all the incoming (gray) Milestones are located within this region with a minimal distance to any of the outgoing Milestones.

The proper selection of the conformational images  $X_1, \dots, X_N$  will be explained in more detail in section 2.3; for now we assume their arbitrary placement. If  $\Delta_i$  were omitted in the above definition ( $\Delta_i = 0$ ), then the set of Milestones  $M_{i \rightarrow j}$  is reduced to the Voronoi tessellation proposed in refs 7 and 8; we refer to this arrangement as Markovian Milestoning with Voronoi Tessellation [MMVT] throughout this paper. In the MMVT arrangement, the Milestone  $M_{i \rightarrow j}$  is equivalent to the Milestone  $M_{j \rightarrow i}$  and the only information they preserve

is the identity of last crossed Milestone, not the direction of such a crossing. (In a private communication, Vandeneijnden disclosed an extension of MMVT to make the Milestones velocity dependent.)

It is important to emphasize that the proposed placements of Milestones is not a tessellation. In accord with the definition of the original Milestoning, a trajectory is identified by the last Milestone that it passes and not by its actual current position. A memory is carried out in time until the trajectory crosses another interface (Milestone).

Trajectories from  $X_i$  to  $X_j$  can be fundamentally different from trajectories from  $X_j$  to  $X_i$ . To exploit this observation it is useful to make the Milestones dependent on the direction. We therefore call Milestones defined according to eq 1 Directional Milestones. The role of the additional flexibility offered by  $\Delta_i$  is to avoid counting rapid transitions between interfaces because of spatial proximity of Milestones. As a result, the Milestones defined by eq 1 depend on more than the coordinates alone. This is consistent with the notion of a Milestone  $M_{i \rightarrow j}$  ( $M_{k \rightarrow j}$ ) as a state of a trajectory that arrives from the region  $X_i$  ( $X_k$ ) to the region of image  $X_j$ . Hence the definition of a Milestone is extended to include information about the previous assignment of the trajectory. If the system is assigned to a region  $X_{i_0}$  at time 0 then by following a trajectory of the system one can deterministically identify the sequence of Milestones the trajectory has passed through  $M_{i_0 \rightarrow i_1}, M_{i_1 \rightarrow i_2}, M_{i_2 \rightarrow i_3}, \dots, M_{i_{K-1} \rightarrow i_K}$ .

**2.2. Calculation of Mean First Passage Times.** In the rest of the manuscript, we will use Roman subscripts to denote image index (as was done in the previous section) and Greek letters to denote Milestones. Consider the mean first passage time (MFPT) from any Milestone  $\alpha$  to a given target Milestone  $\beta$ . We define it as follows: a trajectory is *assigned* to a Milestone  $\alpha$  if the last Milestone it has passed through is  $\alpha$ . *One-step transition* from a Milestone  $\alpha$  to a Milestone  $\beta$  ( $\beta \neq \alpha$ ) is a change of assignment of a trajectory from  $\alpha$  to  $\beta$ . This step is clearly on a coarse Milestoning level and does not mean a single Molecular Dynamics step, which we will call a time-step. If such an event is possible, we say that  $\alpha$  connects to  $\beta$ . Note that by the definition given in eqs 1, if  $\alpha$  connects to  $\beta$ , the second index of  $\alpha$  (e.g.,  $M_{i \rightarrow j}$ ) must be equal to the first index of  $\beta$  ( $M_{j \rightarrow k}$ ). The first hitting point distribution on  $\beta$ ,  $\rho_\beta(p)$ , is the distribution of phase space points (denoted by  $p$ ) at which an equilibrium trajectory passes through  $\beta$  numerous times, while the previous Milestone it passes through was not  $\beta$ . In further discussion, only the relative weight of trajectories that pass through  $\beta$  is important so we can choose to normalize  $\rho_\beta(p)$  such that  $\int \rho_\beta(p) dp = 1$ . We denote by  $\langle \tau_{\alpha\beta}(p) \rangle$  the mean time of all trajectories that start from the phase space point  $p$  in  $\alpha$  and terminate on Milestone  $\beta$  (possibly crossing other Milestones on the way). Integrating the last entity over  $p$ , weighting it by the probability that  $p$  is a phase space point at which an equilibrium trajectory hits  $\beta$  for the first time,  $\int \langle \tau_{\alpha\beta}(p) \rangle \rho_\alpha(p) dp \equiv \langle \tau_{\alpha\beta} \rangle$ , we obtain the MFPT from  $\alpha$  to  $\beta$ .

Let the distribution of one-step transitions from  $\alpha$  to  $\beta$  be  $T_{\alpha\beta}(p, q, t)$ , where  $p$  is the phase space point at which a trajectory starts in  $\alpha$  and  $q$  is the phase space point at which

the trajectory changes its assignment to  $\beta$  after time  $t$ .  $T_{\alpha\beta}(p,q,t)$  is normalized in such a way that if we integrate over  $t$  and  $q$  we get conditional probability of a trajectory reaching  $\beta$  in one step given that it originates from  $p$  in  $\alpha$ :  $\int \int T_{\alpha\beta}(p,q,t) dq dt = P(\beta|\alpha,p)$ , or alternatively  $\sum_{\beta} \int \int T_{\alpha\beta}(p,q,t) dq dt = 1$ . Note that by the definition of trajectory assignment,  $T_{\alpha\alpha}(p,q,t) = 0$  for all  $p$  and  $q$  (since a trajectory cannot change its assignment from  $\alpha$  to  $\alpha$ ).

Assuming that the phase space point  $p(t + dt)$  can be determined from  $p(t)$  only, as is true for most microscopic dynamics (e.g., Newtonian, or Langevin dynamics, but not Generalized Langevin dynamics) we make the following argument: The MFPT from  $\alpha$  to  $\beta$ ,  $\langle\tau_{\alpha\beta}\rangle$ , is defined as the weighted average of termination times of trajectories from  $\alpha$  to  $\beta$ . Each trajectory, starting at  $p$  in  $\alpha$  jumps in one step to some other Milestone  $\gamma$  ( $\gamma \neq \alpha$ ) at phase point  $q$  and then in multiple steps (possibly 0, if  $\gamma = \beta$ ) continues to  $\beta$ . Consider all the trajectories that jump in one-step from  $p$  in  $\alpha$  to  $q$  in  $\gamma$  exactly in time  $t$  and then eventually reach  $\beta$  (in potentially different total time). Since the microscopic dynamics is Markovian we can replace the contribution of these trajectories to  $\langle\tau_{\alpha\beta}\rangle$  by  $(t + \langle\tau_{\gamma\beta}(q)\rangle)$  weighted by sum of the weights of all of them (which is  $\rho_{\alpha}(p)T_{\alpha\gamma}(p,q,t)$ ). By doing this for all possible combinations of  $\gamma$  and  $q$  we get the following equation:

$$\begin{aligned} \langle\tau_{\alpha\beta}\rangle &= \sum_{\gamma} \int \int \int \rho_{\alpha}(p) T_{\alpha\gamma}(p,q,t) (t + \langle\tau_{\gamma\beta}(q)\rangle) dp dq dt \\ &= \sum_{\gamma} \int \rho_{\alpha}(p) \left( \int \int T_{\alpha\gamma}(p,q,t) dt dq \right) dp \\ &+ \sum_{\gamma} \int \langle\tau_{\gamma\beta}(q)\rangle \left( \int \int \rho_{\alpha}(p) T_{\alpha\gamma}(p,q,t) dp dt \right) dq \end{aligned} \quad (2)$$

The first term of the above equation can be reduced as

$$\begin{aligned} \sum_{\gamma} \int \rho_{\alpha}(p) \left( \int \int T_{\alpha\gamma}(p,q,t) dt dq \right) dp &= \sum_{\gamma} \int \rho_{\alpha}(p) \\ &\left( \frac{\int \int T_{\alpha\gamma}(p,q,t) dt dq}{\int \int T_{\alpha\gamma}(p,q,t) dt dq} \int \int T_{\alpha\gamma}(p,q,t) dt dq \right) dp = \\ &\sum_{\gamma} \int \rho_{\alpha}(p) \langle t_{\alpha\gamma}(p) \rangle P(\gamma|\alpha,p) dp = \\ \int \rho_{\alpha}(p) \left( \sum_{\gamma} \langle t_{\alpha\gamma}(p) \rangle P(\gamma|\alpha,p) \right) dp &= \int \rho_{\alpha}(p) \langle t_{\alpha}(p) \rangle dp = \langle t_{\alpha} \rangle \end{aligned} \quad (3)$$

where  $\langle t_{\alpha\gamma}(p) \rangle$ ,  $\langle t_{\alpha}(p) \rangle$ , and  $\langle t_{\alpha} \rangle$  are average times of one-step transitions from  $p \in \alpha$  to  $\gamma$ , from  $p \in \alpha$  to any other Milestone, and from  $\alpha$  to any other Milestone (averaged over  $p$ ), respectively. In the second term of eq 2, the average time from  $q \in \gamma$  to  $\beta$  is weighed by a factor depending on the phase space point  $p \in \alpha$ . To overcome this problem, we use the following assumption: *The distribution at which any Milestone  $\gamma$  is hit does not depend on the Milestone to which the trajectory was assigned before the hit.*

$$\forall \alpha, \gamma: \rho_{\gamma}(q) \propto \int \rho_{\alpha}(p) T_{\alpha\gamma}(p,q,t) dp dt \quad (4)$$

It is easier to illustrate the properties of eq 4 if we consider a one-dimensional arrangement of Milestones in which the

forward and the backward Milestones occupy the same spatial coordinates. Consider a Milestone  $\alpha$  that is pointing forward and is therefore denoted for the clarity of this discussion by  $\alpha+$ . There are two Milestones that initiate trajectories that may terminate at  $\alpha+$ . They are  $(\alpha - 1)+$  and  $(\alpha - 1)-$ . Hence, they occupy the same place in space but have their velocities pointing in the opposite directions. The assumption of eq 4 states that it does not matter if we start at  $(\alpha - 1)+$  or at  $(\alpha - 1)-$ , both Milestones will generate the same hitting point distribution on  $\alpha+$ . If the initial direction of the velocity decorrelates quickly, there should be no difference in the results from Milestone  $(\alpha - 1)+$  and  $(\alpha - 1)-$ . In this case, the assumption formulated in eq 4 will be satisfied. Indeed, we observed empirically in ref 9 that even the usual Milestoning works well when the velocity decorrelates. This empirical finding is now formulated mathematically. In higher dimension, we will also require spatial decorrelation.

The multiplicative factor in the above equation is determined by the fact that if both sides of eq 4 are integrated over  $q$ , the left side equals to 1 and the right side to  $P(\gamma|\alpha)$ ; the conditional probability that if a trajectory changes its assignment from  $\alpha$  it changes to  $\gamma$ . Therefore using the above assumption the second term of eq 2 reduces to  $\sum_{\gamma} P(\gamma|\alpha) \langle\tau_{\gamma\beta}\rangle$ , and we obtain the final form for the MFPT.

$$\langle\tau_{\alpha\beta}\rangle = \langle t_{\alpha} \rangle + \sum_{\gamma} P(\gamma|\alpha) \langle\tau_{\gamma\beta}\rangle \quad (5)$$

The set of eq 5 is supported by boundary conditions  $\langle\tau_{\beta\beta}\rangle = 0$ ,  $\langle t_{\beta} \rangle = 0$ , and  $\forall \gamma P(\gamma|\beta) = 0$ . It is a set of linear equations for all the  $\langle\tau_{\alpha\beta}\rangle$  that can be solved by any standard linear solver. The size of the problem (the number of Milestones) never exceeded a few hundreds in our hands. Equation 5 can be directly generalized for considering more than a single target Milestone (e.g., all incoming interfaces to the folded state of a peptide). Alternative equations equivalent to eq 5 were derived in refs 1 and 4. These equations are independent of the type of microscopic dynamics that we use (e.g., overdamped Langevin or Newtonian). The system of linear eqs 5 relates the overall rate ( $\tau$ 's) with the local kinetics information ( $\langle t_{\alpha} \rangle$  and  $P(\gamma|\alpha)$ ). Milestoning collects this local information in a more effective way than running an ensemble of trajectories from  $\alpha$  to  $\beta$ . On each Milestone  $\alpha$ ,  $N_{\alpha}$  phase space points are sampled from the FHPD  $\rho_{\alpha}$  (see section 2.4 for details). As a second step, each of the sampled phase space points is propagated in time until a connected Milestone is reached. The termination times of these trajectories are typically several orders of magnitude shorter than the overall MFPT of the system. Furthermore the trajectories between Milestones are independent of each other and thus can be run in parallel. For each Milestone  $\gamma$  connected to  $\alpha$ , we record  $N_{\alpha\gamma}$ , the number of trajectories that are initiated on  $\alpha$  and terminated on  $\gamma$ . We also record  $\bar{t}_{\alpha}$ , the mean termination time of all  $N_{\alpha}$  trajectories regardless of their terminal Milestone. The collected information  $\{N_{\alpha\gamma}, \bar{t}_{\alpha}\}$  is used to estimate the required entities for eq 5 as

$$P(\gamma|\alpha) \cong N_{\alpha\gamma}/N_{\alpha} \text{ and } \langle t_{\alpha} \rangle \cong \bar{t}_{\alpha} \quad (6)$$



In practice instead of using eq 6, we employ Bayesian inference on the collected data to calculate the MFPT supported by the data, as well as an estimate of the statistical error due to the finite size of collected data. This procedure is described in detail in Appendix B.

**2.3. Properties of Directional Milestones.** The use of eq 5 for calculating MFPT depends on validity of the assumption expressed in eq 4. It has been shown in<sup>4</sup> that the assumption formulated in eq 4 holds if overdamped Langevin dynamics is used, and the Milestones are chosen as isocommittor surfaces. To our knowledge, there is no efficient algorithm that identifies exact isocommittor surfaces and scales moderately with system size. However, there are other ways of satisfying eq 4. Instead, we base our strategy on selecting Milestones according to eq 1, making sure that Milestones are sufficiently separated to allow for a memory loss of trajectories as outlined in the arguments of ref 1. Consider a pair of connected Milestones  $M_{i \rightarrow j}$ ,  $M_{j \rightarrow k}$  (defined by coordinate images  $X_i$ ,  $X_j$ , and  $X_k$ ). Let  $S_{jk}$  be a hyperplane perpendicular to the line segment  $X_j - X_k$  and passing through its midpoint. From eq 1 that defines  $M_{i \rightarrow j}$  we know that each point on  $M_{i \rightarrow j}$  is closer to  $X_j$  than to  $X_k$ . Thus the Milestone  $M_{i \rightarrow j}$  lies on the  $X_j$ 's side of  $S_{jk}$ . It follows from Lemmas A.1 and A.2 in Appendix A that  $S_{jk}$  and  $M_{j \rightarrow k}$  are parallel,  $M_{j \rightarrow k}$  lies on the  $X_k$ 's side of  $S_{jk}$ , and that  $d(S_{jk}, M_{j \rightarrow k}) = (\Delta_j^2)/(2d(X_j, X_k))$ . Therefore  $d(M_{i \rightarrow j}, M_{j \rightarrow k}) \geq (\Delta_j^2)/(2d(X_j, X_k))$ . This minimal separation of connected Milestones is a property of Directional Milestoning that allows for some velocity relaxation to at least approximately satisfy the assumption described in eq 4. Note that the lower bound for the distance  $d(M_{i \rightarrow j}, M_{j \rightarrow k})$  is a function of distances between the images that we place at will. Minimal separation of any two images places a lower bound on  $\Delta_j$ 's; additionally if one guarantees for each connected pair  $M_{i \rightarrow j}$ ,  $M_{j \rightarrow k}$  that  $d(X_j, X_k)$  is about  $\Delta_j$  then  $d(M_{i \rightarrow j}, M_{j \rightarrow k}) \approx \Delta_j/2$ .

**2.4. Sampling of the First Hitting Point Distribution.** The first step of Milestoning is to sample the initial conditions on each Milestone  $\alpha$  from the first hitting point distribution  $\rho_\alpha(p)$ . An analytical expression for  $\rho_\alpha(p)$  is in general unknown. In ref 4, the authors provided the formula  $\rho_\alpha(x) \propto e^{-\beta V(x)} |\nabla q(x)|$  for the case of overdamped Langevin dynamics with Milestones being placed as isosurfaces of the committor function  $q(x)$ . The last formula includes the gradient of committor function  $\nabla q(x)$  which is difficult to get in high dimensions.

Instead of computing  $\rho_\alpha(p)$  exactly (no exact expression is available for Newtonian dynamic), we approximate it. First, phase space points are sampled from the equilibrium distribution at Milestone  $M_{i \rightarrow j}$ . It can be done either by running an MD simulation constrained to the Milestone,<sup>1,3</sup> or by employing the Umbrella Sampling technique (see appendix C and ref 10). The second step involves filtering each of the sampled phase points to determine those that are indeed first hitting events of  $M_{i \rightarrow j}$ . Exact verification tracks each of the sampled phase space points  $p$  back in time and tests termination on one of the incoming Milestones to the cell  $X_i$  ( $M_{k \rightarrow i}$ ) before the trajectory intersects any of  $M_{i \rightarrow l}$ . (If  $M_{i \rightarrow j}$  itself is crossed before any of  $M_{k \rightarrow i}$ ,  $p$  is not the first hitting event of  $M_{i \rightarrow j}$ , it is at least a second hit of  $M_{i \rightarrow j}$ ; if

$M_{i \rightarrow l}$ ,  $l \neq j$ , is crossed before any of  $M_{k \rightarrow i}$  then the trajectory must have entered the cell of  $X_i$  before reaching  $p$ ; therefore,  $p$  cannot be the first hitting event of  $M_{i \rightarrow j}$ ). Tracking the trajectory back in time to any of the Milestones  $M_{k \rightarrow i}$  is similar in spirit to Transition Interface Sampling,<sup>11–13</sup> (TIS), the difference is that a TIS trajectory is propagated back in time until the reactant or the product state is hit. In DiM, we perform significantly shorter backward verification, applicable only for equilibrium processes. TIS is exact; however, it is more expensive, since in Milestoning we still exploit the use of trajectory fragments. Trajectory fragments are easier to parallelize, and they can lead to implicit long time trajectories, while in TIS, long time individual trajectories need to be computed explicitly.

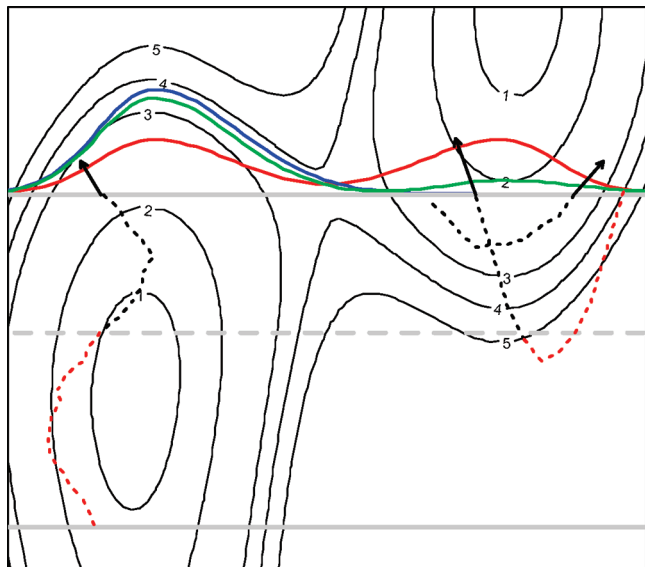
To retain high efficiency we track the trajectory back in time only until it reaches an empirical test boundary that is placed at a distance  $d$  on the  $X_i$ 's side of the target Milestone  $M_{i \rightarrow j}$  ( $d$  being smaller than or equal to the minimal distance to any of  $M_{k \rightarrow i}$  from  $M_{i \rightarrow j}$ ). If the trajectory reaches the checking boundary without recrossing any other Milestone  $M_{i \rightarrow l}$ , we assume that  $p$  is a first hitting event. Otherwise we reject it. The procedure is schematically illustrated on Figure 3.

In principle, we can follow the trajectory back in time until one of the incoming Milestones to  $X_i$  ( $M_{k \rightarrow i}$ ) or any of the outgoing Milestones from  $X_i$  ( $M_{i \rightarrow j}$ ) is hit (a comment by Giovanni Ciccotti). By performing this complete verification the prepared ensemble on each Milestone would be the exact first hitting point distribution. However, the complete verification of each of the sampled phase points roughly doubles the overall computational cost (assuming reasonable acceptance ratio). The result of the more expensive exact verification will be reported elsewhere; in this paper we report results and analysis of the more efficient (but approximate) checking protocol.

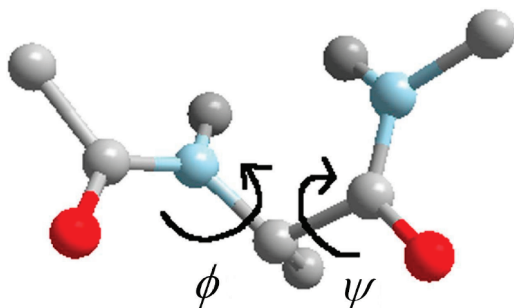
### 3. Applications of Directional Milestoning

**3.1. Alanine Dipeptide Solvated in Water.** To demonstrate an application of Directional Milestoning, we compute MFPT of the transition between  $\alpha$  helix and  $\beta$  sheet conformations in solvated alanine dipeptide (Figure 4). The thermodynamics and kinetics of alanine dipeptide has been investigated in several studies.<sup>1,8,14–16</sup> In aqueous solution, two dihedral angles,  $\phi$  and  $\psi$ , shown in Figure 4, are adequate coarse variables for the dynamics of the peptide. We therefore use a 2-norm distance in the reduced space of  $\phi$  and  $\psi$  as the distance metric in the definition of Milestones (periodicity of the angles was taken into account in the calculation of a distance between two torsion angles).

The new module for Directional Milestoning was created in the program MOIL,<sup>17</sup> and is available at <https://wiki.ices.utexas.edu/clsb/wiki>. The peptide molecule is solvated in a periodic box (20 Å)<sup>3</sup> of 248 TIP3P water molecules. The OPLS force field<sup>18</sup> is used with electrostatics real space cutoff of 9 Å augmented with Particle Mesh Ewald summation. van der Waals interactions are cut at a distance of 8 Å. All calculations were run in NVT ensemble at a temperature of 303 K by employing a weak Andersen thermostat that acts

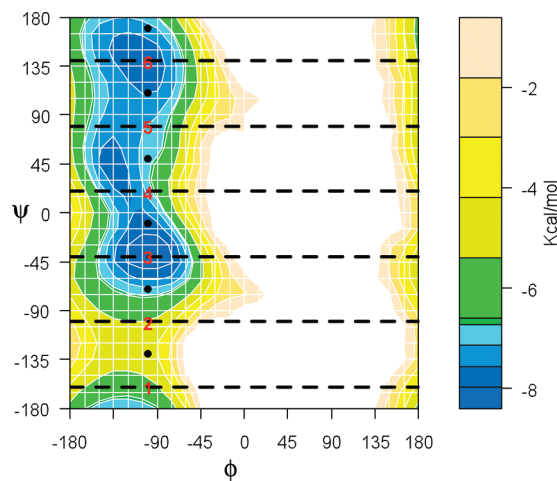


**Figure 3.** Illustration of sampling of the first hitting point distribution of trajectories initiated on the lower gray Milestone and terminating on the top (target) Milestone. The FHPD on the target Milestone (blue) is centered in the left basin, which is different from the equilibrium distribution (red). The FHPD is approximated by sampling phase space points from the equilibrium distribution and following each of them back in time until it hits the target Milestone on which it was initiated (the point is rejected) or the test boundary shown as a dashed gray line (it is accepted). Tracking of three phase space points is shown; the algorithm tracks only the black parts of the trajectories. Two of the points are accepted; one of them, however, is accepted by a mistake. The point is accepted because the test boundary was reached, however if the trajectory were checked further on (the red part) it would have been detected that the trajectory turns back and is not coming from the lower Milestone. Because of these false positive samples, the resulting distribution (green) only approximates  $\rho_{\alpha}(p)$  (blue). As the test boundary approaches the originating Milestone (lower gray), the sampled distribution approaches the true FHPD.



**Figure 4.** Alanine dipeptide.

only on the center-of-mass motion of the water molecules.<sup>19</sup> The probability of velocity resampling was set to  $5 \times 10^{-4}$  per fs. For a water box of this size, an average of 13 water molecules had their velocities resampled in a 100 fs interval. This weak coupling does not change the transition rate obtained from NVE (Newtonian) simulations (with initial conditions sampled from the NVT ensemble). The free energy surface as a function of the two dihedral angles ( $\phi$ ,  $\psi$ ) is shown in Figure 5. It was calculated from statistics of



**Figure 5.** Free energy profile of alanine dipeptide as a function of the two dihedral angles  $\phi$  and  $\psi$ . It was calculated from statistics of a 340 ns long MD simulation. Images for DiM calculations are placed at the positions of the red numbers and for MMVT calculation at the location of the black points. Both algorithms with these placement of images infer the Milestones in the positions of the dashed lines; in DiM, however there are two directional Milestones for each line.

a 340 ns long MD simulation. The white region of the map was not visited by the trajectory. There are two local free energy minima corresponding to an  $\alpha$  helix conformation ( $\phi$ ,  $\psi = -100$ ,  $-40$ ) and to a  $\beta$  sheet conformation ( $\phi$ ,  $\psi = -100$ ,  $140$ ).

The height of the free energy barrier between the two metastable regions at 303 K is less than  $2k_B T$  and the transitions between the metastable states are rapid on the trajectory time scale so the MFPTs can be estimated from straightforward MD simulations directly. We have performed five independent MD simulations of 68 ns. In each of the simulations, more than 1000 transitions between the metastable regions occurred. The MFPT of  $\alpha \rightarrow \beta$  transition is 66.4 ps ( $\pm 2.7$  ps) and that of the opposite transition is 53.8 ps ( $\pm 4.6$  ps). We set up the Milestoning calculation by placing six images in the conformational space in the positions  $\phi_i, \psi_i = -100^\circ, -240^\circ + 60^\circ i$ , ( $i = 1, \dots, 6$ ). The positions of the images were not optimized. They were placed equidistantly in the region of conformation space that is accessible to the molecule. Table 1 shows the results of the Milestoning calculations for this system; it also includes the results of Markovian Milestoning with Voronoi Tessellation method.<sup>7</sup> The MMVT calculation was performed with the same settings as for DiM, with the exception of the image placement; images for MMVT calculation were placed at  $\phi'_i, \psi'_i = -100^\circ, -210^\circ + 60^\circ i$  (for  $i = 1, 2, \dots, 6$ ), so that the Milestones are placed in the same positions as in Directional Milestoning.

Note that the employed dynamics is almost deterministic and thus a trajectory reflected from an interface (procedure required in MMVT) would approximately track itself back in time. Therefore we have slightly modified the MMVT protocol in a way suggested by Vanden-Eijnden in a private communication: instead of reversing the velocities of all the degrees of freedom at a cell interface, only the velocities of peptide atoms are reversed. This modification should not

**Table 1.** Results of the MFPT Calculations on Alanine Dipeptide Solvated in Water with 6 Cells Placed as Shown on Figure 5<sup>a</sup>

method	MFPT [ps], (sd [ps]) $\alpha \rightarrow \beta / \beta \rightarrow \alpha$	total cost [ns]
straightforward MD	<b>66.4</b> (2.7)/ <b>53.8</b> (4.6)	68
DiM, 100 trajectories/Milestone	66.5 (11.1)/39.0 (4.6)	5.0 + 0.6 = 5.6
DiM, 250 trajectories/Milestone	57.7 (5.4)/46.5 (3.6)	12.5 + 1.0 = 13.5
DiM, 500 trajectories/Milestone	61.2 (4.2)/46.8 (2.6)	22.8 + 2.0 = 24.8
DiM, 1000 trajectories/Milestone	57.0 (2.7)/45.2 (1.8)	46.1 + 3.9 = 50.0
DiM, 5000 trajectories/Milestone	59.5 (1.3)/44.2(0.8)	230 + 10.1=240.1
MMVT, 0.4 ns/cell	60.2/43.9	2.4
MMVT, 0.8 ns/cell	57.2/43.7	4.8
MMVT, 1.6 ns/cell	63.2/41.2	9.6
MMVT, 3.4 ns/cell	63.4/53.2	20.4
MMVT, 12 ns/cell	62.4/48.3	72.0

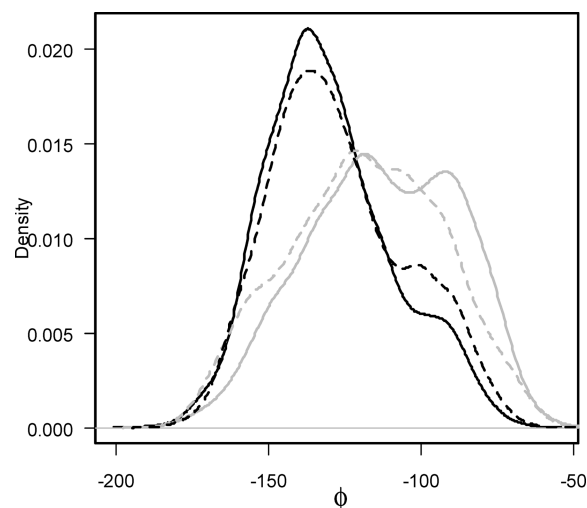
<sup>a</sup> Exact MFPTs were calculated by running five 68 ns long MD trajectories. The standard deviation of predicted MFPT of DiM and MD calculations are given in the brackets. For DiM, standard deviation was calculated from a single execution by using Bayesian inference (details in Appendix B). The total cost for DiM is given as a sum of the simulation time of all trajectories and the simulation time used for preparation of initial distributions.

influence the statistics of observed fluxes through the interfaces since only the peptide degrees of freedom are used in the definition of cell boundaries.

Both methods, DiM and MMVT, perform well in this scenario, though MMVT is more efficient for this simple system. If enough sampling is done, both techniques provide reasonable estimates of MFPTs between the metastable regions, the systematic error is lower for MMVT (6% and 10%) as compared to our method (10% and 18%). Analysis of MMVT on the same system was performed recently.<sup>8</sup> A different force field was employed in ref 8, and the MFPT reported differs by a factor of 2 from our calculations; however the relative error of MMVT for the reported  $\alpha \rightarrow \beta$  transition is about 6%, which is comparable to our result. Results of  $\beta \rightarrow \alpha$  transition were not reported in ref 8. Table 1 shows that MMVT needs about 2–3 times less CPU time compared to DiM to converge. DiM requires more computations in these setting since each interface of MMVT is effectively doubled for the two different directions. Furthermore, additional computation is needed in DiM to sample initial phase space points on each interface. In this one-dimensional setup of Milestones with relatively large separation between Milestones and low free energy barrier, MMVT is more efficient and as accurate as DiM. However, we will show below that with smaller separation between the interfaces, multidimensional arrangement of milestones, and rougher energy landscapes, DiM is better.

Even though previous Milestoning studies calculated accurately MFPTs on alanine dipeptide, memory effects in the system are not negligible. First hitting point distributions (in terms of  $\phi$  angles) for the Milestones  $M_{4 \rightarrow 5}$  and  $M_{6 \rightarrow 5}$  are shown on Figure 6. There is a noticeable difference between distributions of first hitting points on the Milestone  $M_{4 \rightarrow 5}$  and on the Milestone  $M_{6 \rightarrow 5}$ . As shown on the figure, the approximate sampling described in section 2.4 distinguishes the first hitting point distributions arriving from different directions to the region of image  $X_5$  reasonably well.

In Table 2, we examine the use of directional Milestones on this system. The table shows that transitions between the six Milestones (if direction is not part of the description) are not Markovian. If no memory effects were present in the system then the probability of transiting to Milestone  $i + 1$  from Milestone  $i$  would not depend on the Milestone visited before  $i$ , that is the second and the forth columns of



**Figure 6.** Distributions of  $\phi$  angle of the first hitting point conformations of the region of image  $X_5$  (located at  $\psi = 80^\circ$ ): distributions observed in a long MD simulation for conformations arriving to the hypersurface at  $X_5$  from the hypersurface of  $X_4$  (black solid), or from that of  $X_6$  (gray solid). Distributions sampled on the Milestone  $M_{4 \rightarrow 5}$  (black dashed) and the Milestone  $M_{6 \rightarrow 5}$  (gray dashed).

Table 2 would be the same within the error bars. We however see differences of up to 21% (for  $i = 5$ ) or by a factor of up to 2.2 (for  $i = 1$ ). One can see that the values of these relative probabilities estimated by Directional Milestoning (columns 3 and 5 in Table 2) are in good agreement with the true values.

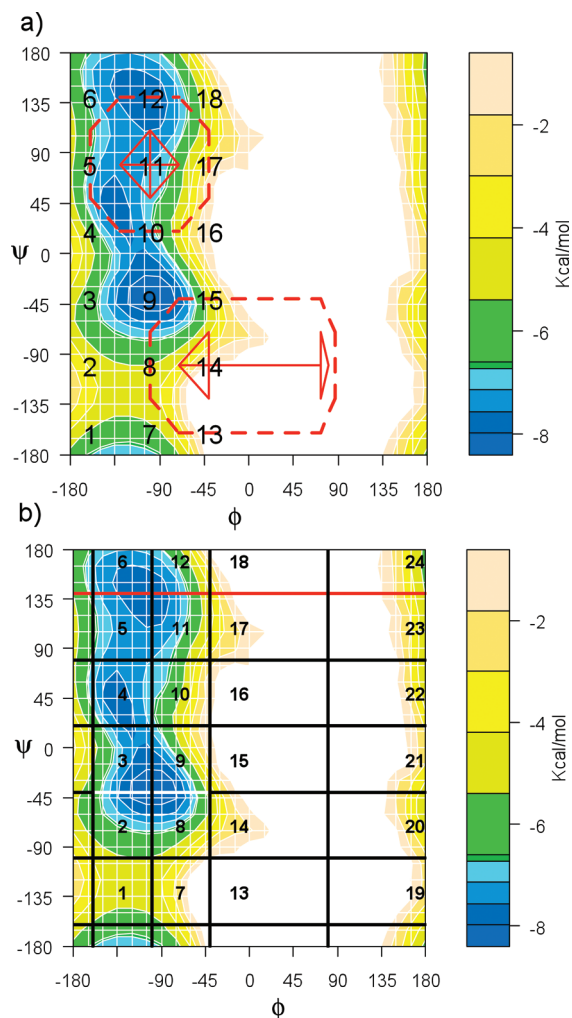
In the second experiment, we examine both methods (DiM and MMVT) on the same system with Milestones in more than one dimension. This experiment is performed to empirically illustrate that placing Milestones in a nonlinear arrangement does not compromise accuracy of DiM calculations. Images are placed in a two-dimensional grid covering the accessible space at the target temperature (conformations with torsional angle  $\phi < 0$ ). For DiM, 18 images are placed in the positions marked 1, ..., 18 on Figure 7a). Each image has 8 incoming Milestones and 8 outgoing Milestones (displayed in solid and dashed on Figure 7a) respectively). We calculated the MFPT from  $M_{12 \rightarrow 11}$  (or  $M_{10 \rightarrow 11}$ ) to the union of  $M_{10 \rightarrow 9}$  and  $M_{8 \rightarrow 9}$  for the  $\beta \rightarrow \alpha$  transition. The MFPTs from these two Milestones differ from each other by about



**Table 2.** Dynamics of the Alanine Dipeptide System Are Not Fully Reducible to a Markov Jump Process between Six Hypersurfaces Shown on Figure 5<sup>a</sup>

$i$	$P(i \rightarrow i+1   i-1 \rightarrow i)$	$N_{M_{i-1} \rightarrow i} M_{i \rightarrow i+1} / N_{M_{i-1} \rightarrow i}$	$P(i \rightarrow i+1   i+1 \rightarrow i)$	$N_{M_{i+1} \rightarrow i} M_{i \rightarrow i+1} / N_{M_{i+1} \rightarrow i}$
1	3.9	3.6	8.6	8.3
2	82.4	84.8	89.4	92.0
3	84.9	88.1	91.0	88.0
4	39.0	37.5	49.0	50.0
5	39.2	41.4	60.6	50.5
6	26.3	32.0	35.0	34.1

<sup>a</sup> The probability of jumping to the Milestone  $i+1$  from the Milestone  $i$  depends on the Milestone visited before  $i$ . Probabilities (from a long MD trajectory) of jumping from  $i$  to  $i+1$  if the Milestone  $i-1$  ( $i+1$ ) was visited before the hypersurface  $i$  are listed in the second (fourth) column. The third and fifth columns list these probabilities as measured by DiM calculation by starting 1000 trajectories from each Milestone. Note that in contrast to DiM, the original Milestoning assumes that  $P(i \rightarrow i+1 | i-1 \rightarrow i) = P(i \rightarrow i+1 | i+1 \rightarrow i)$ .



**Figure 7.** Placement of images on a two-dimensional grid. (a) DiM settings: total of 18 images, located at position of numbers in the plot, are placed in a two-dimensional grid. For two of the images,  $X_{11}$  and  $X_{14}$ , the outgoing (dashed) and incoming (solid) Milestones are shown. (b) Arrangement for MMVT. Twenty-four images are placed in the conformational space so the resulting milestones are in the positions equivalent to DiM. Average of MFPT from the two white Milestones to any of the red Milestones is reported in the results for  $\alpha \rightarrow \beta$  transition.

0.3 ps, and we report their average in Table 3. The opposite transition ( $\alpha \rightarrow \beta$ ) was defined in the equivalent way.

For MMVT, the images were placed in slightly different positions than for DiM (see Figure 7b) such that the

Milestones inferred by the Voronoi Tessellation are in equivalent positions to those used in Directional Milestoning. For the  $\alpha \rightarrow \beta$  transitions, we calculated the MFPT of trajectories starting from the two white Milestones in Figure 7b ( $M_{2 \rightarrow 3}$  and  $M_{8 \rightarrow 9}$ ) and terminating at the union of the red Milestones. MFPT of the transitions from these two starting points differ by less than 0.2 ps so only their average is reported in Table 3. The  $\beta \rightarrow \alpha$  calculation was performed in the equivalent way (from the two central Milestones in the  $\beta$  sheet conformation ( $\psi = 140^\circ$ ) to the union of all the Milestones with  $\psi = -40^\circ$ ).

The results of both methods are listed in Table 3. The accuracy of Directional Milestoning is not compromised by multidimensionality; hence DiM works well for higher dimensions or higher connectivity of Milestones. The relative error of the MMVT method increased to 33% (31%). We think that this is mainly because of the corners between Milestones in the MMVT arrangement that cause rapid termination times between nearby Milestones and unwanted correlations between touching Milestones. Evidence of this can be seen in Figure 8.

**3.2. Alanine Dipeptide in Vacuum.** In vacuum, there are two stable conformers  $C_{7eq}$  and  $C_{ax}$  of alanine dipeptide (Figure 9). The state  $C_{7eq}$  is further split into two substates denoted by  $C_{7eq}$  and  $C'_{7eq}$  (located at  $X_{26}$  in Figure 9b) separated by a small barrier. We calculate the MFPT of transition from  $C_{7eq}$  to  $C_{ax}$  at two different temperatures, 400 and 350 K, using Langevin dynamics. This is performed by calculating MFPT starting from each of the incoming Milestones to  $C_{7eq}$  region (green on Figure 9) and considering union of the incoming Milestones to the region  $C_{ax}$  (red on Figure 9) as the final state. The MFPT is not sensitive to exact identity of the starting Milestone (variation of less than 2%) therefore an average MFPT from all green Milestones is considered. The friction constant of Langevin dynamics was set to  $30 \text{ ps}^{-1}$ .

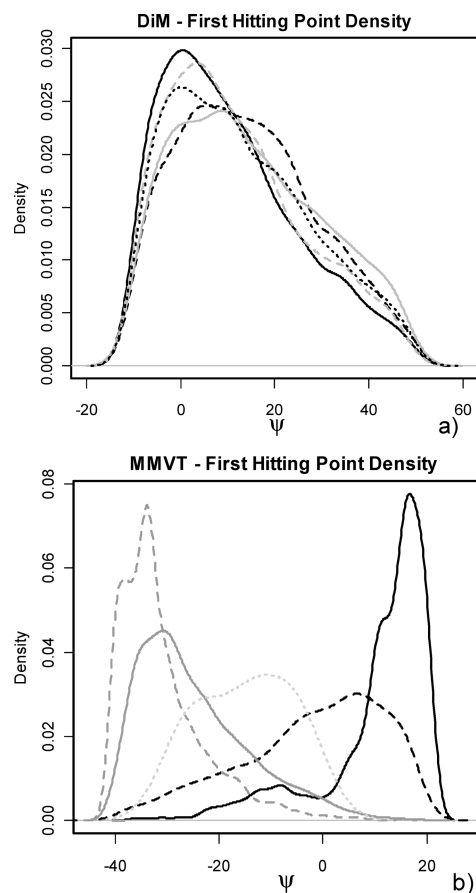
**3.2.1. Image and Cell Generation.** The images were generated by the following expansion. We start with the set of images  $S = \{X_1, X_2\}$ , where  $X_1$  is a conformation located at  $C_{ax}$  and  $X_2$  at  $C_{7eq}$ . Then we iteratively pick an image  $X$  from the set  $S$  and “expand” it: We launch trajectories starting from  $X$  with randomly initiated velocities and run each of these trajectories until it departs at least a prespecified distance  $\delta$  from  $X$ . Then we cluster the set of end points of these trajectories to existing images in  $S$  and potentially add new images to the set  $S$  if there are end points that are farther



**Table 3.** Results of the MFPT Calculations on Alanine Dipeptide Solvated in Water with 18 Cells Placed As on Figure 7a<sup>a</sup>

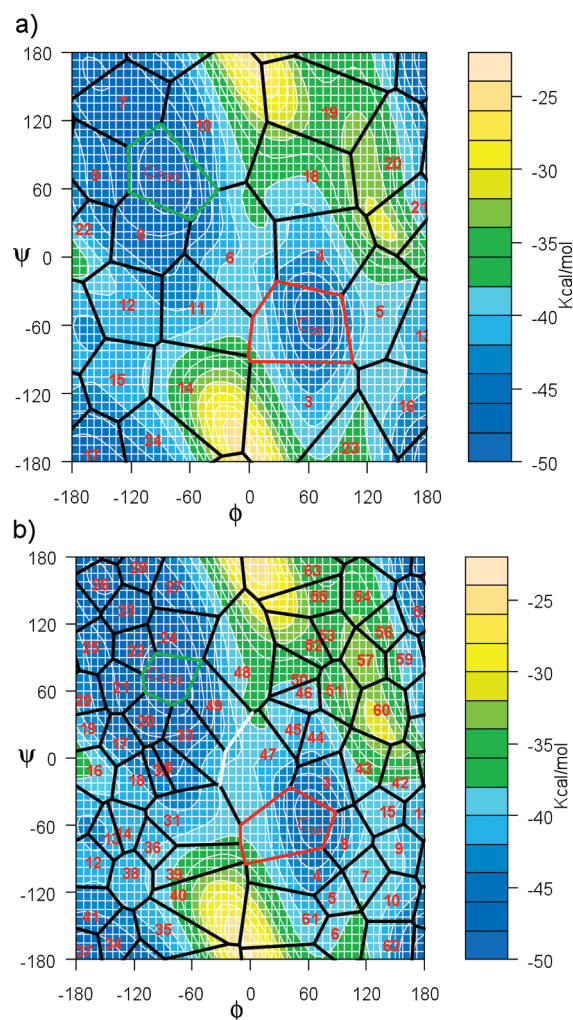
method	MFPT [ps], (sd [ps]) $\alpha \rightarrow \beta/\beta \rightarrow \alpha$	total cost [ns]
straightforward MD	<b>66.4</b> (2.7)/ <b>53.8</b> (4.6)	68
DiM, 100 trajectories/Milestone	68.2 (10.0)/56.9 (8.9)	10.0 + 2.6 = 12.6
DiM, 300 trajectories/Milestone	63.5 (4.9)/56.6 (4.1)	31.1 + 8.7 = 39.8
DiM, 1000 trajectories/Milestone	62.8 (2.5)/53.2 (1.6)	103 + 26 = 129
DiM, 2000 trajectories/Milestone	65.7 (1.6)/52.2 (1.1)	207 + 52 = 259
MMVT, 5 ns/cell	48.6/37.0	120
MMVT, 10 ns/cell	44.3/37.1	240

<sup>a</sup> Standard deviations are in the brackets. Total cost for DiM is given as a sum of the simulation time of all trajectories and the simulation time used for preparation of the initial ensemble on each Milestone.



**Figure 8.** First hitting point distributions. (a) For DiM, distribution of  $\psi$  torsional angle of conformations arriving to the Milestone  $M_{4 \rightarrow 10}$  from the Milestones  $M_{9 \rightarrow 4}$  (black, solid),  $M_{11 \rightarrow 4}$  (black, dashed),  $M_{10 \rightarrow 4}$  (gray, solid),  $M_{3 \rightarrow 4}$  (gray, dashed), and  $M_{5 \rightarrow 4}$  (black, dotted). (b) For MMVT, distribution of  $\psi$  torsional angle of conformations arriving to the Milestone  $M_{3 \rightarrow 9}$  from the Milestones  $M_{10 \rightarrow 9}$  (black, solid),  $M_{4 \rightarrow 3}$  (black, dashed),  $M_{8 \rightarrow 9}$  (gray, solid),  $M_{2 \rightarrow 3}$  (gray, dashed), and  $M_{15 \rightarrow 9}$  (gray, dotted). Note that the target Milestone  $M_{3 \rightarrow 9}$  in the MMVT arrangement is in the same position as  $M_{4 \rightarrow 10}$  in the DiM arrangement, only shifted by  $-30^\circ$  in  $\psi$  direction.

than  $\delta$  from all images of  $S$ . We repeat this process until no new images are generated, that is, we have tried launching trajectories from all images in  $S$  and all end coordinates are in  $S$ . There are three parameters in this algorithm: (i) the distance cutoff  $\delta$ , (ii) the number of expanding trajectories  $N_e$ , and (iii) the clustering algorithm employed. For alanine dipeptide, we have used expectation-maximization as a clustering algorithm,<sup>20</sup> with  $N_e$  set to 400 and two different values of  $\delta$ ,  $\delta_1 = 0.6 \text{ \AA}$  and  $\delta_2 = 0.4 \text{ \AA}$ . The root mean



**Figure 9.** Adiabatic  $\phi$ ,  $\psi$  energy map. The energy is minimized, while constraining the  $\phi$  and  $\psi$  dihedrals to specified values. Placement of (a) 24 images, (b) 63 images in the conformational space based on the algorithm described in section 3.2.1 is shown. Also displayed is the Voronoi Tessellation based on the periodic Euclidean metric in the reduced space of  $\phi$  and  $\psi$  torsions.

squared distance after optimal overlap<sup>21</sup> (rmsd) is the distance metric (the rmsd between  $X_1$  and  $X_2$  is  $1.25 \text{ \AA}$ ) for the purposes of clustering as well as the distance function in the definition of Milestones (1).

**3.2.2. Results for Alanine Dipeptide in Vacuum.** By using different values for  $\delta$  we obtained sets of images of size 24 (for  $\delta_1$ ) and 63 (for  $\delta_2$ ); both are shown on Figure 9. The tessellations shown in black in this figure are only approximate since they are based on the Euclidean distance in

**Table 4.** Results of the MFPT Calculations on Alanine Dipeptide in Vacuum with 24 Cells Placed as on Figure 9a at Temperature 400 K<sup>a</sup>

method	MFPT [ns]	total cost [ $\mu$ s]
straightforward MD at $T = 400$ K	375 (16)	150
DiM, 500 trajectories/Milestone	630 (299)	$0.13 + 0.09 = 0.22$
DiM, 1K trajectories/Milestone	217(103)	$0.26 + 0.18 = 0.46$
DiM, 3K trajectories/Milestone	306 (76)	$0.78 + 0.47 = 1.25$
DiM, 10K trajectories/Milestone	344 (37)	$2.6 + 1.6 = 4.2$
DiM, 20K trajectories/Milestone	387 (34)	$5.2 + 3.1 = 8.3$
DiM, 30K trajectories/Milestone	352 (31)	$7.8 + 4.7 = 12.5$
MMVT, 10 ns/cell	135	0.24
MMVT, 20 ns/cell	289	0.48
MMVT, 40 ns/cell	322	0.96
MMVT, 60 ns/cell	359	1.5
MMVT, 130 ns/cell	351	3.1
MMVT, 400 ns/cell	336	9.6

<sup>a</sup> Standard deviations are in the brackets. Estimation of the exact MFPT was performed by launching five groups of 400 trajectories from  $C_{7eq}$  state and running them until  $C_{ax}$  state is reached (the MFPT reported in the table is calculated as the MFPT of all 2000 trajectories; the error is estimated by standard deviation of MFPTs calculated from each of the five groups). Total cost for DiM is given as a sum of the simulation time of all trajectories and the simulation time used for preparation of the initial ensemble on each Milestone.

**Table 5.** Results of the MFPT Calculations on Alanine Dipeptide in Vacuum with Cells Placed as on Figure 9a, b at Temperature 350 K<sup>a</sup>

method	MFPT [ $\mu$ s]	total cost [ $\mu$ s]
straightforward MD at $T = 350$ K	2.05 (0.3)	410
DiM, 5K trajectories/Milestone	2.78 (0.65)	$2.3 + 1.4 = 3.7$
DiM, 10K trajectories/Milestone	1.74 (0.40)	$4.7 + 2.8 = 7.5$
DiM, 20K trajectories/Milestone	1.75 (0.33)	$9.4 + 5.6 = 15.0$
DiM, 60K trajectories/Milestone	1.77 (0.20)	$28 + 16.8 = 44.8$
MMVT, 24 cells, 2.00 $\mu$ s/cell	69.7	48
MMVT, 63 cells, 0.75 $\mu$ s/cell	3798	47
MMVT, 63 cells, 2.25 $\mu$ s/cell	855	142

<sup>a</sup> DiM was performed with 24 cells, MMVT in two different settings: 24 and 63 cells. Standard deviations are in the brackets. Estimation of the exact MFPT was performed by launching five groups of 200 trajectories from  $C_{7eq}$  state and running them until  $C_{ax}$  state is reached. Standard deviation and average of the MFPT calculated from each group are reported in the table. Total cost for DiM is given as a sum of the simulation time of all trajectories and the simulation time used for preparation of the initial ensemble on each Milestone.

( $\phi, \psi$ ) space, where the real interfaces (Milestones) are defined using the rmsd distance. The MFPT of the transitions between the metastable conformations are significantly longer than those in the solvated peptide because of higher free energy barriers. Tables 4 and 5 summarize the results of the Milestoning calculations in this system. At the high temperature (400 K) both methods, DiM and MMVT, predict accurate MFPT from  $C_{7eq}$  to  $C_{ax}$  (with systematic error of about 10%). MMVT needs to run about 1.5  $\mu$ s MD simulations to obtain converged results, while DiM requires about 2.5  $\mu$ s. Both of them provide significant speed up against straightforward MD simulation, even though a rough estimate of MFPT of the  $C_{7eq}$  to  $C_{ax}$  transition can be obtained by running about 11 independent MD simulations (equivalent to 4  $\mu$ s of the total simulation time); however, both MMVT and DiM can be trivially parallelized to thousands of CPUs, shortening the actual time to perform the calculation.

When the temperature is lowered to 350 K (see Table 5) the  $C_{7eq}$  to  $C_{ax}$  transition is slower with MFPT of about 2.0  $\mu$ s. As listed in Table 5, Directional Milestoning calculates the MFPT with systematic error of about 15% with as few as 7.5  $\mu$ s of total simulation time. That is significant speedup compared to straightforward MD since DiM can be easily parallelized on thousands of processors. MMVT fails to calculate the MFPT accurately. The main reason for this failure is poor statistics. An important difference between DiM and MMVT is that DiM allocates computational

resources to each Milestone, where MMVT allocates the computational resources to a cell. If a transition between two specific interfaces in a cell is needed to describe the reaction and the transition is significantly less likely than transitions between other interfaces of the cell, then sampling this transition using MMVT is inefficient. A simple realization of this effect is the existence of a barrier in the middle of the cell. In that case MMVT trajectory is likely to be confined to a one minimum, to collide with the same interface many times (hits that do not count for the statistics) and to record only a few successful transitions to the other minimum. In contrast, DiM launches a large number of short trajectories. These trajectories terminate quickly, and contribute to the statistics.

In DiM, sampling is done (extensively) at the interfaces, so the probability of observing a transition between interfaces of interest is greatly enhanced, since at least one end of the transitional event is sampled extensively. A potential problem in DiM is a large number of interfaces that may make sampling expensive. To avoid sampling irrelevant interfaces (at a given temperature) trajectories are initiated at few initial interfaces and only interfaces that are hit at least once during the DiM calculation are sampled and launched. We stop the DiM calculation when the process converges (i.e., no new interfaces besides those already sampled are reached).

For the MMVT calculation with 24 images, many cells cover a relatively large part of the conformational space with

a rough energy landscape (see for example cell  $X_6$  on Figure 9a). This arrangement may cause poor statistics for those regions since the trajectories spend most of their time in low free energy regions, rarely visiting interfaces higher in free energy. To increase the probability of having a double hit at the two desired surfaces, we run the same calculation with 63 images as well. But even when 63 images are used, the allocation of computational resources is highly unbalanced. For example, we consider the frequency of hitting the interfaces  $49 \rightarrow 47$  and  $33 \rightarrow 47$  (displayed in white on Figure 9b) that are both important for the overall MFPT. In both, 49 and 33 cells, confined simulations of total time of  $2.25 \mu\text{s}$  hit a cell boundary more than  $2 \times 10^7$  times. However, the interface  $33 \rightarrow 47$  is hit only 17 times and the interface  $49 \rightarrow 47$  only 7 times. In contrast DiM allocates an equal number of starting trajectories to each of the Milestones and transitions from Milestones located near the transition states are sampled as well as other Milestones. We have not experimented with any selection criterions for allocation of computational resources to different cells (in MMVT) or to different Milestones (in DiM) but both methods may benefit from selective allocation of resources to “important regions” of conformational space.

#### 4. Discussions and Conclusions

In this paper, we proposed a method to compute dynamics in high dimensions called Directional Milestoning. We have shown that the mean first passage times between Milestones can be calculated accurately given that the distribution at which a Milestone is hit does not depend on the previously assigned Milestone (the assumption formulated in eq 4). Directional Milestoning arranges dividing hypersurfaces in a special way, aiming to satisfy the above assumption: (i) Milestones in DiM are made directional, so the local progress of the reaction (going from the region of  $X_i$  to  $X_j$  as opposed to being at the interface between  $X_i$  and  $X_j$ ) is made part of the description, (ii) the arrangement of Milestones guarantees a lower bound on spatial separation of any connected pair of Milestones so trajectories initiated on a Milestone have space and time to “lose memory” before terminating on a different Milestone.

The algorithm, while based on the trajectory fragments of Milestoning, is a step in the direction of Transition Interface Sampling<sup>11–13</sup> (TIS) and the Forward Flux Sampling (FFS) methods<sup>22,23</sup> compared to the original Milestoning. Here we use some trajectory tracking. The main difference between these methods and Directional Milestoning is that TIS and FFS are tracking trajectories all the way back to the reactant state. This tracking has the advantage of not relying on any assumption about the initial ensemble on an interface like is done in Milestoning. On the other hand, sampling of trajectories in TIS and FFS is computationally more expensive than in Milestoning because every attempted trajectory in these methods is tracked back to the reactant state where in (Directional) Milestoning a trajectory is tracked only until it reaches a different Milestone. Computations of trajectory fragments can be done in

Milestoning in a massively parallel way. The PPTIS method uses a conceptually similar approach of trajectory fragments.<sup>24</sup>

An important distinction of Directional Milestoning compared to TIS, FFS, and the original Milestoning is that it allows for arbitrary arrangement of Milestones in conformational space, not necessarily following a linear arrangement along an order parameter or a reaction coordinate. A similar (arbitrary) arrangement of interfaces is used in the MMVT method,<sup>7</sup> nonequilibrium umbrella sampling method,<sup>25,26</sup> and Trajectory Parallelization and Tilting method.<sup>27</sup> The last two techniques are using short trajectories in cells and balance the fluxes between cells. Recently the nonequilibrium umbrella sampling<sup>26</sup> was illustrated to be more efficient than FFS.<sup>28</sup> The Weighted Ensemble approach was also shown recently to work without a reaction coordinate.<sup>29</sup>

We have compared DiM with MMVT and showed that the performance of MMVT (in terms of effectiveness and correctness) is comparable to that of DiM in some of the examples, but that the correctness and/or effectiveness of MMVT can be compromised in systems with high free energy barriers, or in cells with two interfaces that are hard to reach. Another problem for straightforward implementation of MMVT is the existence of corners between Milestones along more than one dimension that contribute to termination times that are too short. So while DiM is in general somewhat slower than MMVT it provides reliable results more consistently, including cases in which MMVT fails.

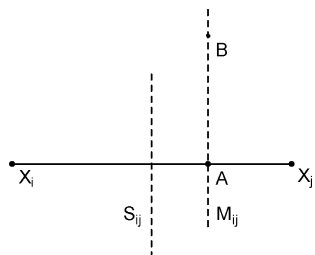
We also would like to comment on the similarities (and the differences) of our approach to the Markov State Model (MSM; for a recent study see ref 30). In the applications of MSM that we are aware of, long to very long Molecular Dynamics trajectories at normal conditions are used to estimate transition times and population of different cells. MMVT and DiM are designed to avoid such long trajectories (at the cost of approximate matching of probability densities at the interfaces). Once a sample of conformational space is available (which can be done in numerous ways, reaction path calculations, replica exchange simulations, or high temperature trajectories) only very short Molecular Dynamics trajectories are required to estimate the local kinetics. These short trajectories can be trivially parallelized providing profound computational saving compared to straightforward Molecular Dynamics simulations. While significant progress has been made in parallelizing a single trajectory,<sup>31</sup> overhead still remains and special hardware that is frequently used is more expensive to buy and to maintain.

**Acknowledgment.** This research was supported in part by NIH grant GM059796 and NSF grant 0833162 to RE.

#### Appendix A: Lemmas Regarding the Milestones Geometry

**Lemma A.1.** Let  $X_i$  and  $X_j$  be two images in conformation space such that  $M_{i \rightarrow j}$  exists. Let  $A$  be an intersection of the line segment  $X_i X_j$  with  $M_{i \rightarrow j}$ . Then a point  $B$  on the hyperplane





perpendicular to  $X_i X_j$  and passing through  $A$  belongs to  $M_{i-j}$  iff  $\forall k d(X_k, B) \geq d(X_j, B)$ .

**Proof of Lemma A.1.** From def 1 of  $M_{i-j}$

$$d(X_i, A)^2 - d(X_j, A)^2 = \Delta_i^2$$

By using the Pythagoras theorem for triangles  $X_i A B$  and  $X_j A B$ .

$$\begin{aligned} d(X_i, B)^2 - d(X_j, B)^2 &= (d(X_i, A)^2 - d(A, B)^2) - (d(X_j, A)^2 - d(A, B)^2) \\ &= d(X_i, A)^2 - d(X_j, A)^2 = \Delta_i^2 \end{aligned}$$

As a consequence of Lemma 1,  $M_{i-j}$  is a hyperplane segment perpendicular to  $X_i X_j$ .

**Lemma 2.** Let  $S_{ij}$  be the hyperplane perpendicular to the line segment  $X_i X_j$  and passing through its midpoint. Then  $d(S_{ij}, M_{i-j}) = (\Delta_i^2)/(2d(X_i, X_j))$ .

**Proof of Lemma 2.** Since both  $S_{ij}$  and  $M_{i-j}$  are perpendicular to  $X_i X_j$ , the distance  $d(S_{ij}, M_{i-j})$  is equal to the distance of the  $X_i X_j$  midpoint,  $P_{ij}$ , and the intersect of  $M_{i-j}$  with  $X_i X_j$ ,  $A$ . Thus

$$\begin{aligned} (d(X_i, P_{ij}) + d(S_{ij}, M_{i-j}))^2 - (d(X_j, P_{ij}) - d(S_{ij}, M_{i-j}))^2 &= \Delta_i^2 \\ d(S_{ij}, M_{i-j}) &= \frac{\Delta_i^2}{4d(X_i, P_{ij})} = \frac{\Delta_i^2}{2d(X_i, X_j)} \end{aligned}$$

## Appendix B: Statistical Reasoning

We describe an estimate of the statistical error of a milestone calculation from a single set of collected data using Bayesian reasoning. As shown in section 22, eq 5, repeated here as B.1

$$\langle \tau_{\alpha\beta} \rangle = \langle t_\alpha \rangle + \sum_\gamma P(\gamma|\alpha) \langle \tau_{\gamma\beta} \rangle \quad (\text{B.1})$$

relates MFPTs  $\langle \tau_{\alpha\beta} \rangle$  and local kinetics entities ( $\langle t_\alpha \rangle$  and  $P(\gamma|\alpha)$ ). Milestoning aims to estimate  $\langle t_\alpha \rangle$  and  $P(\gamma|\alpha)$  by launching  $N_\alpha$  trajectories from each Milestone  $\alpha$ .  $N_{\alpha\gamma}$  of them terminate on the Milestone  $\gamma$  and the mean incubation time (time to termination) of all  $N_\alpha$  trajectories is  $\bar{t}_\alpha$ . In Bayesian inference, a statistical model of the transitions among Milestones is needed. We closely follow and extend notation used in the analysis of Markovian Milestoning with Voronoi Tessellations; for more details consult.<sup>7</sup> The same kinetic formulas (with different notation) are also available from ref 9. We assume a continuous Markov jump process between the Milestones controlled by a transition matrix  $Q$  defined in the following way: let the probability distribution of the system over all the Milestones be  $\rho = (\rho_1, \dots, \rho_N)$ , where  $\rho_\alpha$  is the probability that the system is assigned to a

Milestone  $\alpha$ . Under continuous Markov jump process,  $\rho$  behaves as

$$\dot{\rho} = \rho Q \quad (\text{B.2})$$

For transition matrix  $Q$ , by definition  $q_{\alpha\alpha} = -\sum_{\beta \neq \alpha} q_{\alpha\beta}$  and it can be shown by simple algebra that  $P(\beta|\alpha) = q_{\beta\alpha} / \sum_{\gamma \neq \alpha} q_{\gamma\alpha}$  and  $\langle t_\alpha \rangle = 1 / \sum_{\gamma \neq \alpha} q_{\gamma\alpha}$  (for derivation see, for example, refs 1, 2, and 7). By plugging the last three identities to the linear system (B.1) it reduces to

$$Q' \langle \tau \rangle = -1 \quad (\text{B.3})$$

where  $\langle \tau \rangle$  is the row vector ( $\langle \tau_{1\beta} \rangle, \dots, \langle \tau_{\beta-1\beta} \rangle, \langle \tau_{\beta+1\beta} \rangle, \dots, \langle \tau_{N\beta} \rangle$ )<sup>T</sup> and  $Q'$  is a  $(N-1) \times (N-1)$  matrix created from  $Q$  by skipping the row and the column related to the Milestone  $\beta$ . To infer  $\langle \tau \rangle$  from the collected data,  $\{N_{\alpha\gamma}, \bar{t}_\alpha\}$ , using eq B.3, a relation between  $\{N_{\alpha\gamma}, \bar{t}_\alpha\}$  and  $Q'$  is needed. Following the derivations from ref 7, for a system ruled by B.2, the probability of staying in a state  $\alpha$  for time  $t$  and then jumping to a state  $\beta$  in the time interval  $\langle t, t + dt \rangle$  is  $e^{-\sum_{\gamma \neq \alpha} q_{\alpha\gamma} t} q_{\alpha\beta} dt$ . Using this equality, the likelihood of observing the collected data,  $L(\{N_{\alpha\gamma}, \bar{t}_\alpha\} | Q)$ , is

$$L(\{N_{\alpha\gamma}, \bar{t}_\alpha\} | Q) = \prod_\alpha \prod_{\gamma \neq \alpha} q_{\alpha\gamma}^{N_{\alpha\gamma}} e^{-q_{\alpha\gamma} N_{\alpha\gamma} \bar{t}_\alpha} \quad (\text{B.4})$$

By using the Bayes' rule the likelihood that the true transition matrix is  $Q$  given the collected data,  $L(Q | \{N_{\alpha\gamma}, \bar{t}_\alpha\})$ , is

$$L(Q | \{N_{\alpha\gamma}, \bar{t}_\alpha\}) \propto \prod_\alpha \prod_{\gamma \neq \alpha} q_{\alpha\gamma}^{N_{\alpha\gamma}} e^{-q_{\alpha\gamma} N_{\alpha\gamma} \bar{t}_\alpha} P(Q) \quad (\text{B.5})$$

where  $P(Q)$  is the prior probability distribution of  $Q$  without seeing any data (typically this is set to uniform if we do not have any prior knowledge about the system). Equation B.5 is typically used in maximum likelihood estimators, for example, one estimates unknown entity  $Q$  with  $Q^*$ , the matrix that maximizes likelihood  $L(Q | \{N_{\alpha\gamma}, \bar{t}_\alpha\})$ . In this particular case,  $Q^*$  has form  $q_{\alpha\gamma}^* = N_{\alpha\gamma} / [N_{\alpha\gamma} \bar{t}_\alpha]$ , which is in agreement with estimators given in eq 6 in the main text. Instead of using purely  $Q^*$  for calculations of MFPTs, we can examine whole distribution of transition matrices according to eq B.5 and understand what is the distribution of MFPTs consistent with the data collected. Therefore we typically sample a number of (typically 300) transition matrices from distribution B.5 and look at the variance of MFPTs predicted by them. If standard deviation of MFPTs is large, it suggests that more data about the system shall be collected. We report standard deviation obtained by this algorithm in the results of section 3.

## Appendix C: Sampling Equilibrium Distribution on a Milestone Using Umbrella Sampling

As described in Section 2.4 the equilibrium ensemble from a Milestone  $M_{i-j}$  is used to sample the first hitting point distribution on the Milestone  $M_{i-j}$ . The Milestone  $M_{i-j}$  is defined in eq 1 as  $M_{i-j} \equiv \{X | d(X, X_i)^2 = d(X, X_j)^2 + \Delta_i^2 \text{ and } \forall k d(X, X_j) \leq d(X, X_k)\}$ , where  $\{X_1, \dots, X_K\}$  is a set of images



in the conformational space. In practice, we work with the following approximation of  $M_{i \rightarrow j}$ :

$$\begin{aligned} d_{ij}(X) &\equiv d(X, X_j)^2 - d(X, X_i)^2 + \Delta_i^2 \\ M'_{i \rightarrow j} &\equiv \{X | \forall k, d(X, X_k) \geq d(X, X_j) \wedge -\lambda \leq d_{ij}(X) \leq 0\} \end{aligned} \quad (\text{C.1})$$

Clearly as  $\lambda \rightarrow 0$ ,  $M'_{i \rightarrow j}$  converges to  $M_{i \rightarrow j}$ . We have used  $\lambda = 0.5^\circ$  or  $\lambda = 0.01 \text{ \AA}$  for the calculations on alanine dipeptide.

To sample conformations in  $M'_{i \rightarrow j}$  from equilibrium distribution the following Umbrella Sampling protocol is employed. We run NVT trajectory of the system (using Andersen thermostat) with a modified potential function  $U$  and examine a conformation every few steps (every 100–400 fs for examples described in this paper). If an examined conformation belongs to  $M'_{i \rightarrow j}$  it is saved; otherwise it is discarded. If conformation is saved, corresponding velocities are sampled from Boltzmann distribution. The potential function  $U$  is modified to bias the system toward the region  $M'_{i \rightarrow j}$  in the following way:

$$\begin{aligned} U'(X) &= U(X) + U_{ij}^1(X) + U_{ij}^2(X) \\ U_{ij}^1(X) &= \begin{cases} K_1 d_{ij}(X)^2 & \text{if } d_{ij}(X) > 0 \\ K_1 (d_{ij}(X) - \lambda)^2 & \text{if } d_{ij}(X) < -\lambda \\ 0 & \text{otherwise} \end{cases} \\ U_{ij}^2(X) &= \begin{cases} K_2 (d(X, X_k) - d(X, X_j))^2 & \text{if } d(X, X_k) < d(X, X_j) \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

By definition for  $X \in M'_{i \rightarrow j}$ ,  $U'(X) = U(X)$  and therefore saved points from  $M'_{i \rightarrow j}$  are sampled with the true equilibrium probabilities. If on the other hand NVT trajectory of the system is outside of the region  $M'_{i \rightarrow j}$ , the terms  $U_{ij}^1$  and  $U_{ij}^2$  force the system to return back to  $M'_{i \rightarrow j}$ , the strength of this bias is controlled by force constants  $K_1$  and  $K_2$  (both are set to  $10^3 \text{ kcal mol}^{-1} \text{ rad}^{-2}$  or  $10^4 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$  for alanine dipeptide system).

## References

- West, A. M. A.; Elber, R.; Shalloway, D. *J. Chem. Phys.* **2007**, *126*, 145104.
- Shalloway, D.; Faradjian, A. K. *J. Chem. Phys.* **2006**, *124*, 054112.
- Faradjian, A. K.; Elber, R. *J. Chem. Phys.* **2004**, *120*, 10880.
- Vanden-Eijnden, E.; Venturoli, M.; Ciccotti, G.; Elber, R. *J. Chem. Phys.* **2008**, *129*, 174102.
- Elber, R. *Biophys. J.* **2007**, *92*, L85.
- Kuczera, K.; Jas, G. S.; Elber, R. *J. Phys. Chem. A* **2009**, *113*, 7461.
- Vanden-Eijnden, E.; Venturoli, M. *J. Chem. Phys.* **2009**, *130*, 194101.
- Maragliano, L.; Vanden-Eijnden, E.; Roux, B. *J. Chem. Theory Comput.* **2009**, *5*, 2589.
- West, A. M. A.; Elber, R.; Shalloway, D. *J. Chem. Phys.* **2007**, *126*.
- Torrie, G. M.; Valleau, J. P. *J. Comput. Phys.* **1977**, *23*, 187.
- van Erp, T. S.; Bolhuis, P. G. *J. Comput. Phys.* **2005**, *205*, 157.
- Moroni, D.; van Erp, T. S.; Bolhuis, P. G. *Phys. A* **2004**, *340*, 395.
- Moroni, D.; Bolhuis, P. G.; van Erp, T. S. *J. Chem. Phys.* **2004**, *120*, 4055.
- Ren, W.; Vanden-Eijnden, E.; Maragakis, P.; E, W. *J. Chem. Phys.* **2005**, *123*, 134109.
- Maragliano, L.; Vanden-Eijnden, E. *J. Chem. Phys.* **2008**, *128*, 184110.
- Ensing, B.; De Vivo, M.; Liu, Z.; Moore, P.; Klein, M. L. *Acc. Chem. Res.* **2005**, *39*, 73.
- Elber, R.; Roitberg, A.; Simmerling, C.; Goldstein, R.; Li, H.; Verkhivker, G.; Keasar, C.; Zhang, J.; Ulitsky, A. *Comput. Phys. Commun.* **1995**, *91*, 159.
- Jorgensen, W. L.; Tirado-Rives, J. *J. Am. Chem. Soc.* **2002**, *110*, 1657.
- Juraszek, J.; Bolhuis, P. G. *Biophys. J.* **2008**, *95*, 4246.
- Hartley, H. *Biometrics* **1958**, *14*, 174.
- Kabsch, W. *Acta Crystallogr., Sect. A* **1976**, *32*, 922.
- Valeriani, C.; Allen, R. J.; Morelli, M. J.; Frenkel, D.; ten Wolde, P. R. *J. Chem. Phys.* **2007**, *127*, 114109.
- Allen, R. J.; Frenkel, D.; ten Wolde, P. R. *J. Chem. Phys.* **2006**, *124*, 024102.
- Moroni, D.; Bolhuis, P. G.; van Erp, T. S. *J. Chem. Phys.* **2004**, *120*, 4055.
- Warmflash, A.; Bhimalapuram, P.; Dinner, A. R. *J. Chem. Phys.* **2007**, *127*, 154112.
- Dickson, A.; Warmflash, A.; Dinner, A. R. *J. Chem. Phys.* **2009**, *130*, 074104.
- Vanden-Eijnden, E.; Venturoli, M. *J. Chem. Phys.* **2009**, *131*, 044120.
- Allen, R. J.; Frenkel, D.; ten Wolde, P. R. *J. Chem. Phys.* **2006**, *124*, 17.
- Zhang, B. W.; Jasnow, D.; Zuckerman, D. M. *J. Chem. Phys.* **2010**, *132*.
- Noe, F.; Schutte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 19011.
- Shaw, D. E.; Deneroff, M. M.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; Chao, J. C.; Eastwood, M. P.; Gagliardo, J.; Grossman, J. P.; Ho, C. R.; Ierardi, D. J.; Kolossvary, I.; Klepeis, J. L.; Layman, T.; McLeavey, C.; Moraes, M. A.; Mueller, R.; Priest, E. C.; Shan, Y. B.; Spengler, J.; Theobald, M.; Towles, B.; Wang, S. C. *Commun. ACM* **2008**, *51*, 91.

## Conjugated Molecules Described by a One-Dimensional Dirac Equation

Matthias Ernzerhof\* and Francois Goyer

Département de Chimie, Université de Montréal, C. P. 6128 Succursale A,  
Montréal, Québec H3C 3J7, Canada

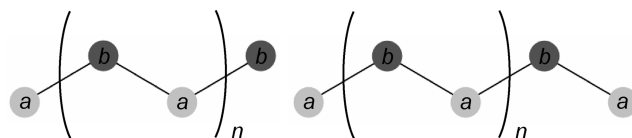
Received January 3, 2010

**Abstract:** Starting from the Hückel Hamiltonian of conjugated hydrocarbon chains (ethylene, allyl radical, butadiene, pentadienyl radical, hexatriene, etc.), we perform a simple unitary transformation and obtain a Dirac matrix Hamiltonian. Thus already small molecules are described exactly in terms of a discrete Dirac equation, the continuum limit of which yields a one-dimensional Dirac Hamiltonian. Augmenting this Hamiltonian with specially adapted boundary conditions, we find that all the orbitals of the unsaturated hydrocarbon chains are reproduced by the continuous Dirac equation. However, only orbital energies close to the highest occupied molecular orbital/lowest unoccupied molecular orbital energy are accurately predicted by the Dirac equation. Since it is known that a continuous Dirac equation describes the electronic structure of graphene around the Fermi energy, our findings answer the question to what extent this peculiar electronic structure is already developed in small molecules containing a delocalized  $\pi$ -electron system. We illustrate how the electronic structure of small polyenes carries over to a certain class of rectangular graphene sheets and eventually to graphene itself. Thus the peculiar electronic structure of graphene extends to a large degree to the smallest unsaturated molecule (ethylene).

### Introduction

Presently, graphene is at the center of numerous investigations (for recent reviews see, e.g., refs 1–5). The remarkable features of graphene, such as its conductance properties,<sup>3</sup> can be attributed to its peculiar electronic structure. In the simplest Hückel description,<sup>6</sup> accounting for nearest-neighbor coupling, the dispersion relation close to the Fermi energy is linear, and this energy region can be described<sup>7,8</sup> by a two-dimensional (2D) Dirac equation of massless fermions. Recently, one-dimensional (1D) graphene-based systems have become accessible to experiments<sup>2,9</sup> as well as zero-dimensional (“molecular”) ones.<sup>2,10,11</sup>

The question that we strive to answer in the present work is whether the electronic structure of graphene has precursors in the molecular domain. Here we refer to the unsaturated hydrocarbon chains defined in Figure 1 as polyenes. Within the Hückel model (nearest-neighbor coupling), we show that these polyenes are described by a 1D, discrete Dirac equation



**Figure 1.** Schematic representation of the unsaturated hydrocarbons investigated. Even- (left) as well as odd-numbered chains are considered. The artificial subdivision into sublattices of *a*- and *b*-type carbon atoms is indicated as well.

of massless ( $m = 0$ ) particles. The continuum limit of this discrete equation yields a 1D Dirac equation that reproduces the Hückel orbitals exactly (up to a gauge transformation and a normalization factor). The orbital energies of orbitals close to the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) are well represented as well by the continuous equation. The continuous, 1D Dirac equation ( $m = 0$ ) contains a Pauli spin matrix:

$$\sigma = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \quad (1)$$

which multiplies the derivative operator  $d = d/dx$ , i.e.:

\* Corresponding author. E-mail: Matthias.Ernzerhof@UMontreal.ca.

$$-iv_F\sigma d\phi = \varepsilon\phi \quad (2)$$

$\phi$  is a two-component pseudospinor:

$$\phi = \begin{pmatrix} \phi_u \\ \phi_d \end{pmatrix} \quad (3)$$

and  $v_F$  the Fermi velocity. As we will show below, this equation, which is often referred to as Weyl equation,<sup>12</sup> is the continuous limit of the transformed Hückel Hamiltonian of linear polyenes. The fact that these molecules are described by a Dirac equation offers a new perspective for their electronic structure. A recent article<sup>13</sup> dealing with infinite, linear carbon chains arrives at an equivalent Dirac equation for the description of the system close to the Fermi energy.

A question that arises in this context is whether the electronic structure of the polyenes can be somehow related to the electronic structure of graphene or finite parts thereof. We provide a positive answer to this question by examining rectangular pieces of graphene. For a particular class of graphene rectangles, the electronic structure near the Fermi energy is essentially 1D and identical to the electronic structure of polyenes.

#### Polyene Described in Terms of a 1D Dirac Equation.

Now we show how the Dirac equation, described in the introduction, emerges as the continuum limit of the Hückel matrix for polyene. To this end, we start from a matrix of unspecified dimension. We regroup the atoms into two bipartite sublattices (cf. Figure 1) such that every other atom belongs to the sublattice of  $a$  atoms, and the remaining ones form the  $b$  sublattice. To emphasize this division, we use the variable  $a$  for the diagonal matrix element of the atoms belonging to the  $a$  sublattice. Similarly,  $b$  denotes the diagonal element of the atom in the  $b$  sublattice, even though the numerical value of both these parameters is zero:

$$\mathbf{H} = \begin{pmatrix} \ddots & & & & & & & & & & \\ & a & t & & & & & & & & \\ & t & b & t & & & & & & & \\ & & t & a & t & & & & & & \\ & & & t & b & t & & & & & \\ & & & & t & a & t & & & & \\ & & & & & t & b & & & & \\ & & & & & & & \ddots & & & \end{pmatrix} \quad (4)$$

where  $t$  is the hopping parameter. An appropriate permutation of the basis functions separates the atoms of the  $a$  and  $b$  lattice. The resulting  $\tilde{\mathbf{H}}$  reads

$$\tilde{\mathbf{H}} = \begin{pmatrix} \ddots & & & & & & & & & & \\ & a & & t & t & & & & & & \\ & & a & & t & t & & & & & \\ & & & a & & t & t & & & & \\ & & & & \ddots & & & & & & \\ \ddots & t & & & & b & & & & & \\ & t & t & & & & b & & & & \\ & & t & t & & & & b & & & \\ & & & \ddots & & & & & \ddots & & \end{pmatrix} \quad (5)$$

$\tilde{\mathbf{H}}$  can be converted into the discrete version of the Dirac Hamiltonian in eq 2 through application of a gauge transformation  $\mathbf{G}$ . This transformation consists of multiplying the coefficient of every second atom of the  $a$  lattice as well as  $b$  lattice with  $-1$ . If we denote the initial coefficient vector by  $\mathbf{C}$ , then the new, transformed coefficient vector will be given by

$$\mathbf{Q} = \mathbf{G}\mathbf{C} \quad (6)$$

$\mathbf{Q}$  represents the envelope function of  $\mathbf{C}$ . For orbitals with energies close to the HOMO and LUMO, the short-range variations in  $\mathbf{C}$  are eliminated by  $\mathbf{G}$ , and only the long-range variations are retained in  $\mathbf{Q}$ . This point is illustrated in the orbital plots provided below. A corresponding modification of the Hamiltonian matrix, compensating for the transformation of the wave function, yields the matrix operator  $\mathbf{D}$ :

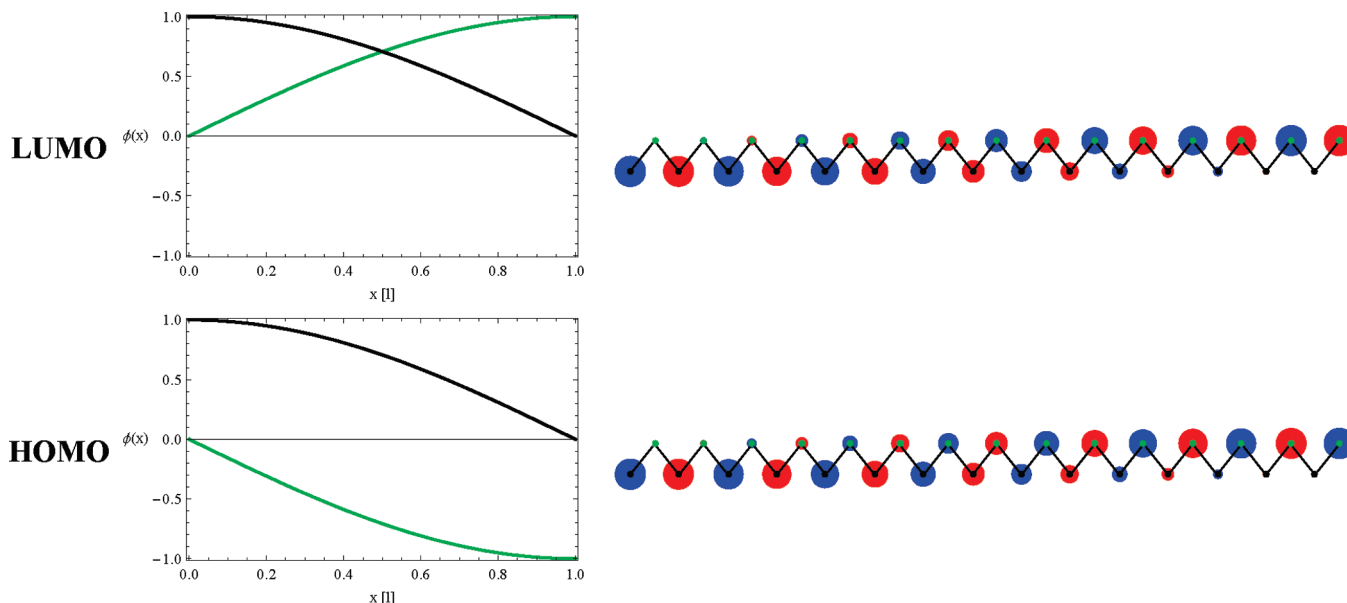
$$\mathbf{D} = \mathbf{G}\mathbf{H}\mathbf{G}^{-1} = \begin{pmatrix} \ddots & & & & & & & & & & \\ & a & & -t & t & & & & & & \\ & & a & & -t & t & & & & & \\ & & & a & & -t & t & & & & \\ & \ddots & & & \ddots & & & & & & \\ \ddots & -t & & & b & & & & & & \\ & t & -t & & & b & & & & & \\ & & t & -t & & & b & & & & \\ & & & \ddots & & & & \ddots & & & \\ & & & & & & & & \ddots & & \end{pmatrix} \quad (7)$$

Keeping in mind that  $v_F = -2t$ ,  $\mathbf{D}$  can be identified as the discretized version of  $-iv_F\sigma d/dx$  appearing in eq 2. The solutions of the continuous Dirac equation (eq 2) are now discussed in some detail and compared to the eigenfunctions obtained with the Hückel approach, or equivalently, to the eigenfunctions of  $\mathbf{D}$ .

**Solutions for Particular Systems: Introducing Boundary Conditions.** The eigenfunctions of the continuous Dirac equation (eq 2) and their respective eigenvalues are of the form:

$$\phi^\pm = \frac{1}{\sqrt{2}}e^{ikx} \begin{pmatrix} 1 \\ \pm i \end{pmatrix}, \quad \varepsilon = \pm v_F k \quad (8)$$

The wave vector  $k$  in this equation is a real variable. Similar to the 2D case, the eigenfunctions exhibit helicity.<sup>4</sup> This means that solutions describing forward-moving particles have a defined pseudospin orientation  $s^+ = \begin{pmatrix} 1 \\ +i \end{pmatrix}$ , whereas backward-traveling particles have the opposite pseudo spin  $s^- = \begin{pmatrix} 1 \\ -i \end{pmatrix}$ . To arrive at this result, one can calculate the probability current density  $j$  to obtain,  $j = v_F(\phi^*\sigma\phi - \phi\sigma\phi^*)$ , and note that  $\sigma s^+ = s^+$  and  $\sigma s^- = -s^-$ . The equation for  $j$  also shows that even if the wave vector in eq 8 is proportional to the energy, the particle velocity is constant and equal to  $v_F$ . In relativistic quantum mechanics this behavior corresponds to massless particles that move with the speed of light. Here we are particularly interested in finding the wave functions of finite systems, and the periodic solutions of eq 2 are not a suitable starting point. Real solutions  $\phi^s$  and  $\phi^c$  that are linear combinations of the complex ones ( $\phi^+$  and  $\phi^-$ ) are appropriate



**Figure 2.** Massless Dirac particle in a box. The  $k = -(\pi/2)/l$ , (where  $n = -1$ ) and  $k = (\pi/2)/l$  (with  $n = 0$ ) wave functions are displayed on the left ( $\phi_u$  is represented by the green and  $\phi_d$  by the black curve) and compared to the HOMO (lower orbital) and LUMO (upper orbital) of a polyene chain with 30 atoms (i.e.,  $N = \text{even}$ ). The continuous solutions are accurate representations of the orbital envelope functions.

$$\phi^s = \begin{pmatrix} \sin(kx) \\ \cos(kx) \end{pmatrix} \quad (9)$$

$$\phi^c = \begin{pmatrix} \cos(kx) \\ -\sin(kx) \end{pmatrix} \quad (10)$$

These states are not pure pseudospin states anymore. In the continuum limit, the equivalent of a polyene would be a finite box terminated to the right and left by infinite potential barriers. Apparently, the appropriate boundary conditions are that the wave function vanishes at the borders (at  $x = 0$  and  $x = l$ ) of the box, i.e.,  $\phi(x = 0) = \phi(x = l) = 0$ . Naively, one would apply this condition to the two components of the pseudospinor. Here, however, we arrive at the continuous model as a limit of discrete systems, and the boundary conditions are also defined through a limiting procedure. Thus, each of the two components represents the wave function on a set of points the two of which are complementary. Therefore, we have to distinguish between two cases: (i) the one where each of the sets contains one end point, implying that the number of atoms ( $N$ ) in the chain is even, and (ii) the case where one set includes both points 0 and  $l$ , which would correspond to an odd number of atoms in the chain. Starting with (i), we suppose that the domain of the up component  $\phi_u$  in the wave function  $\phi = (\phi_u, \phi_d)$  starts at the left boundary of the box and terminates before the right boundary is reached. Similarly, the domain of points for the down component ( $\phi_d$ ) does not contain the boundary point on the left but the one on the right. This choice excludes eq 10, and only eq 9 yields solutions that vanish at the boundary if

$$k = \frac{(n + (1/2))\pi}{l}, \quad n = \dots, -2, -1, 0, 1, 2, \dots \quad (11)$$

To illustrate these solutions, we choose  $n = -1$  and 0 and plot the two components of the wave function (Figure

2). For comparison, we also plot the HOMO and LUMO of a  $N = 30$  polyene chain. Clearly, the solutions of the continuous Dirac equation yield an accurate model for the envelope function of the HOMO (corresponding to  $n = -1$ ) and the LUMO orbital (corresponding to  $n = 0$ ). To better establish the connection between the solutions of the discrete equation and the ones of the continuous equation, we note that an  $N$ -atom chain represents a box that has been discretized with  $N + 2$  points. Two of these points, one on both the left- and the right-hand side, lie on the boundary of the box where the wave function vanishes. Correspondingly, the  $N$  remaining points (associated with the  $N$  atoms) are all inside of the box and have nonvanishing coefficients in general.

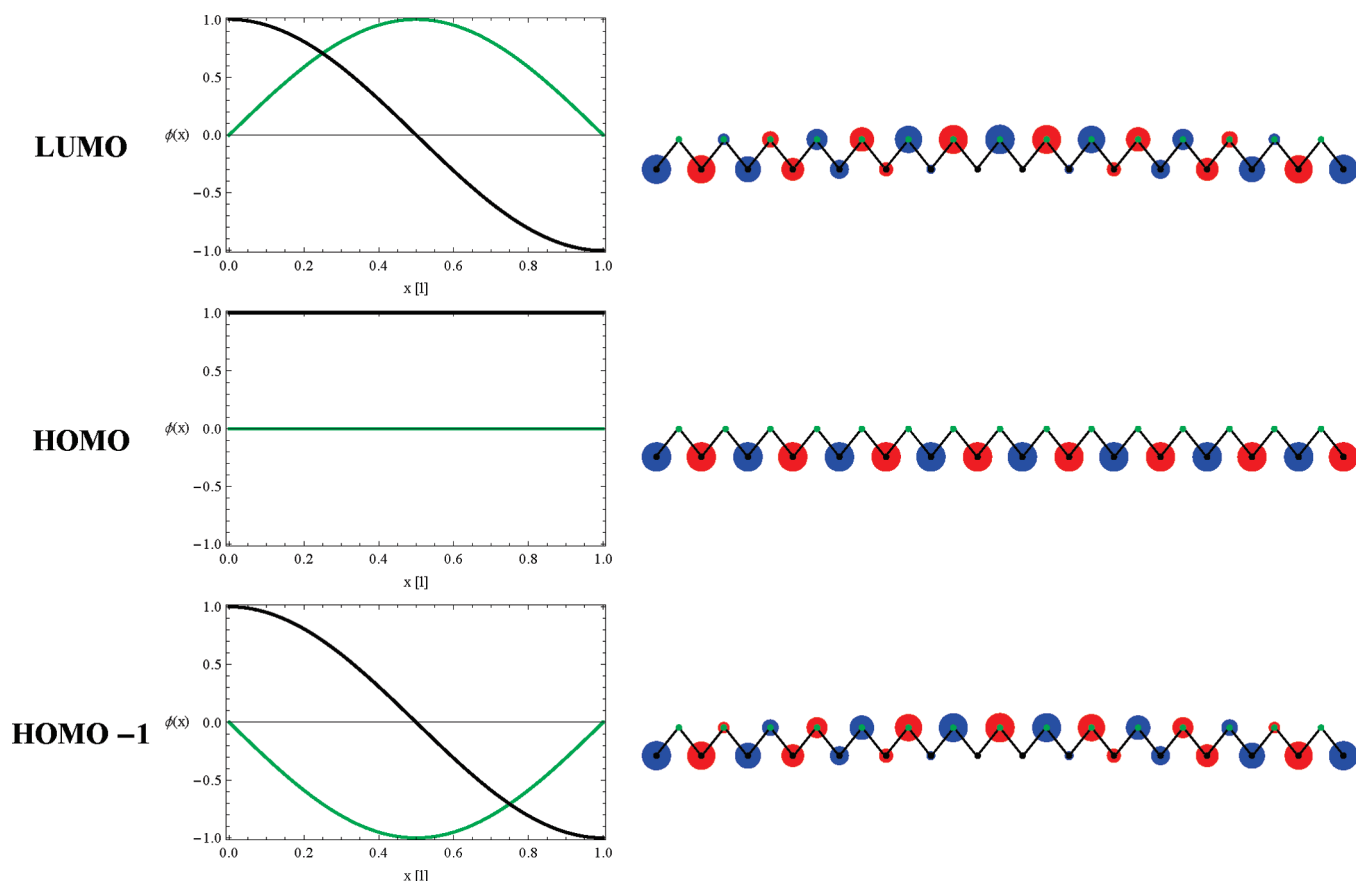
Next we consider the case (ii) where the domain of  $\phi_u$  contains both end points. This would correspond to a carbon chain with an odd number of atoms. Since the boundary condition is only relevant for one of the two discrete sets of points, i.e., the one  $\phi_u$  is defined on, only this component has to satisfy the boundary conditions and thus defines the admissible values of  $k$ :

$$k = \frac{n\pi}{l}, \quad n = \dots, -2, -1, 0, 1, 2, \dots \quad (12)$$

In this case (see Figure 3), the continuous model also reproduces the envelope of the displayed Hückel orbitals.

Another 1D system of interest is a ring where the wave function satisfies periodic boundary conditions. The solutions of the continuous Dirac equation can only be compared to the orbitals of a circular molecule, whose number of atoms is a multiple of 4. This condition arises because the transformation  $\mathbf{G}$  (eq 6) imposes that, on each sublattice, every second coefficient is multiplied by  $-1$ . For this transformation to be unique on a ring, it has to contain a number of atoms that is a multiple of 4. Both solutions of





**Figure 3.** Massless Dirac particle in a box. The left part illustrates the solutions of the continuous Dirac equation with  $k = -\pi/l$ , 0, and  $\pi/l$  (corresponding to  $n = -1, 0, 1$ ), from the bottom up, respectively;  $\phi_u$  is represented by the green and  $\phi_d$  by the black curve. For comparison, the corresponding orbitals of polyene with 31 atoms are also shown. The singly occupied orbital is in the middle, while its energetic neighbors are shown below and above. Again, the envelope function of the orbitals is reproduced by the continuous solutions.

the continuous Dirac equation,  $\phi^s$  and  $\phi^c$ , satisfy the periodic boundary conditions if

$$k = n \frac{2\pi}{l}, \quad n = \dots, -2, -1, 0, 1, 2, \dots \quad (13)$$

As in the nonperiodic case, the solutions of the Dirac equation correspond to the envelope functions of the Hückel orbitals. Alternatively, the complex solutions  $\phi^-$  and  $\phi^+$  can also be employed together with eq 13 to describe the system. In this case the solutions are helical, a property shared with the massless fermions found in infinite graphene. Cyclic molecules with  $4n$  electrons are anti-aromatic, thus for finite systems, helicity and anti-aromaticity appear to coincide.

**Small Polyenes.** Now we turn to the question of how long a polyene has to be to be well represented by the continuum model. To address this issue, we compare the dispersion relation of the continuous Dirac equation to the Hückel energy distribution of polyenes. The latter distribution is known analytically:

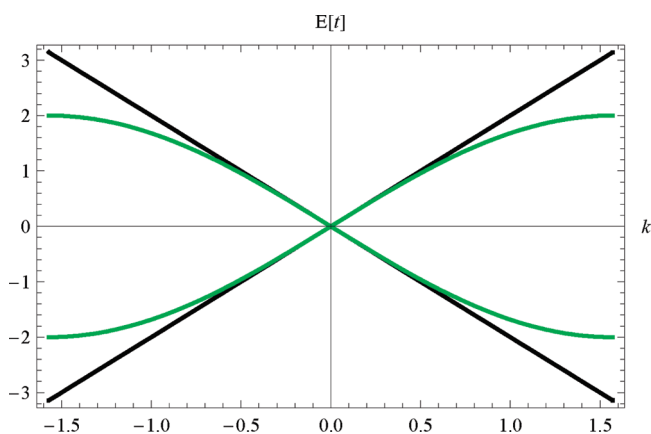
$$E_n = 2t \cos\left(\frac{\pi n}{N+1}\right), \quad n = 1, 2, \dots, N \quad (14)$$

We introduce a wave vector  $k$  in eq 14 in the canonical way. Changes in the wave vector are  $(\pi\Delta n)/(N+1)$ , as the index of the states changes by  $\Delta n$ . Furthermore, the origin of the wave vector is shifted by  $-\pi/2$  to account for the

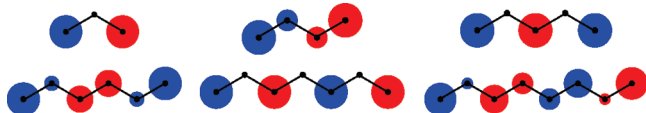
gauge transformation. The resulting dispersion relation:

$$E(k) = -2t \sin(k), \quad k \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \quad (15)$$

is compared (Figure 4) to the corresponding relation of the continuous Dirac equation. For small values of  $|k|$ , the



**Figure 4.** Energy dispersion relation of finite polyenes (green curves) compared to the dispersion relation of the continuous Dirac equation (black lines). Note that the argument  $k$  in eq 15 has been replaced by  $-k$  to obtain the second green graph. This has been done to mimic the degeneracy that is present in the infinite or periodic system but absent in the nonperiodic, finite one.



**Figure 5.** HOMOs of  $N$ -atom polyenes ( $N = 3, \dots, 8$ ). The odd-numbered chains all show identically structured HOMOs having a vanishing contribution on one sublattice. Similarly, the even-numbered chains have HOMOs, whose envelope describes a quarter wave on each of the sublattices.

dispersion relations are almost identical, and they differ in the large- $|k|$  regime, where the Hückel spectrum exhibits the known quadratic behavior.

A related question is how similar the Hückel orbitals are to the solutions of the Dirac equation? This question has a strikingly simple answer. For all values of  $N$  they are, up to the gauge transformation  $\mathbf{G}$  and a normalization factor, identical. More precisely, if we take a solution of the continuous Dirac equation (subject to the proper boundary conditions) and represent it on an equidistant grid of  $N + 2$  points, then the resulting orbital is related to an orbital of a polyene with  $N$  atoms by a gauge transformation. This can be readily verified by acting with  $\mathbf{D}$  on  $\mathbf{Q}$ , where  $\mathbf{Q}$  is a grid representation of  $\phi$ , defined in eq 9. Then components, such as  $t \sin(k(n + 2)) - t \sin(kn)$  and  $t \cos(k(n + 2)) - t \cos(kn)$ , of the resulting vector are examined. By exploiting trigonometric identities, these expressions can be converted to

$$\begin{aligned} t \sin(kn) - t \sin(k(n + 2)) &= -2t \cos(k(n + 1)) \sin(k) \\ &= E(k) \cos(k(n + 1)) \end{aligned} \quad (16)$$

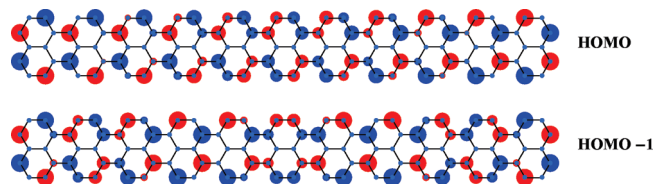
and

$$\begin{aligned} t \cos(k(n + 2)) - t \cos(kn) &= -2t \sin(k(n + 1)) \sin(k) \\ &= E(k) \sin(k(n + 1)) \end{aligned} \quad (17)$$

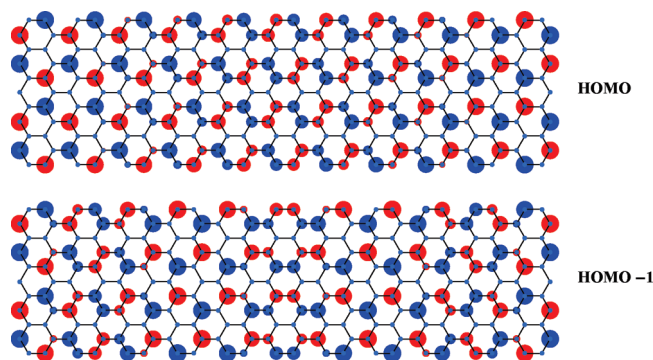
showing that the discrete Dirac equation is satisfied. This finding is analogous to the known<sup>14</sup> fact that a discretization of the free-electron model yields the exact Hückel orbitals.

In Figure 5, we provide the Hückel HOMOs of a number of polyenes with  $N = 3, \dots, 8$ . As can be guessed and verified analytically, they are identical to the solutions of the continuous Dirac equation modulo a gauge transformation.

**Relation to Finite Graphene Ribbons.** The electronic structure of the polyenes discussed above carries over to a surprising degree to finite graphene ribbons<sup>15–17</sup> of a certain width. In detail, we consider a symmetric, rectangular piece of graphene with two armchair edges (along the longitudinal direction) and two zigzag edges. The length of the ribbons is assumed to be large compared to their width. Recently, the edge states that appear in finite graphene ribbons and other finite shapes have been investigated extensively, among these studies are refs 18–27. These edge states are typically found along zigzag borders. Electronic structure calculations taking electron interaction into account show that the edge states can result in spin polarization since they are preferentially occupied with electrons of a particular spin orientation.<sup>18,26</sup> By keeping the width of the ribbon small



**Figure 6.** Shown are the HOMO and the HOMO-1 of a graphene rectangle which is two benzene rings wide. Both orbitals exhibit a nodal plane in longitudinal direction which effectively separates the system into two polyene chains. These polyene chains in turn exhibit envelope functions corresponding to solutions of the continuous Dirac equation.



**Figure 7.** Orbitals of a graphene rectangle which is five benzene rings wide. Omitting states localized at the zigzag boundaries (i.e., edge states discussed in the text), the HOMO and the HOMO-1 are displayed in the figure. Effective polyene chains are clearly visible in both orbitals.

compared to its length, we reduce the number of edge states and the associated complexity.

The narrowest ribbon, conforming to the above listed restrictions, is shown in Figure 6. It exhibits no edge states, and we recover the electronic structure already discussed for polyene. In the examples considered, the transverse degree of freedom is effectively frozen since excitations in transverse direction require too much energy. Furthermore, there is no binding or antibinding in the transverse direction since the states of interest have an energy close to 0. As one might suspect, the pattern observed in Figure 6 repeats itself in ribbons, whose width is given by  $3n + 2$  benzene rings. An example is shown in Figure 7, where  $n = 1$ . This ribbon exhibits four edge states with energies at or close to 0. These states are localized at the zigzag edges, and here we neglect their impact on the electronic structure, which is reasonable if the length of the ribbon is small compared to its width. Infinite armchair graphene ribbons with a width of  $3n + 2$  benzene rings are known to be metallic in the Hückel model (see, e.g. refs 17, 28–31 and references therein). In view of our discussion this is not surprising since, in the limit of an infinite ribbon length, the 1D Dirac equation, describing the dispersion relation around the Fermi energy, yields a vanishing band gap. It should be mentioned, however, that more sophisticated density functional theory calculations<sup>32,33</sup> yield small band gaps for armchair ribbons.

Armchair ribbons other than the ones discussed here exhibit, in general, a coupling between the longitudinal polyene chains, which results in a band gap even at the

Hückel level. Furthermore, we recall that, except for the narrowest armchair ribbons, edge states play a role and alter the dispersion relation close to the Fermi energy. Increasing the length of the ribbon renders these surface effects less and less important. Detailed computational studies of the transition of the electronic structure from finite ribbons to graphene are provided in refs 18 and 22.

## Summary and Conclusion

Maybe the most important result obtained here is that the Hückel Hamiltonian can be rigorously transformed into a discrete Dirac Hamiltonian. Even the continuous limit of the discrete Hamiltonian, yielding the 1D Dirac equation, results in the exact polyene orbitals. The boundary conditions, accompanying the continuous Dirac operator, play a crucial role. They are somewhat involved since they are defined by taking the limit of the conditions applied to discrete systems. Straight forward application of particle-in-a-box boundary conditions does not yield a solution for the Dirac equation of massless particles.

For graphene, the electronic structure close to the Fermi energy is accurately described by the Dirac equation. This result is arrived at (see, e.g., ref 4) by linearizing a wave vector dependent Schrödinger equation that is subjected to periodic boundary conditions. Adding to these findings, here we consider finite polyene chains and show that a surprisingly simple transformation turns the Hückel matrix of a finite polyene into a discrete Dirac Hamiltonian.

Neglecting edge states (see the discussion in the previous section), we find that graphene rectangles that have a width (given in units of benzene rings) of  $3n + 2$  exhibit no coupling between longitudinal and transverse degrees of freedom in orbitals with energies close to 0. These systems can be described as effective polyene chains by the 1D Dirac equation. For armchair graphene ribbons of infinite length, it is known that, at the Hückel level,  $3n + 2$  ribbons have a vanishing band gap.<sup>17</sup> This appears to be a simple consequence of the structure of the wave function revealed here. Increasing the length and width of the  $3n + 2$  ribbon, we eventually arrive at graphene where what we refer to as the longitudinal direction is now one of three equivalent orientations, since graphene has a six-fold rotational symmetry.

Graphene is regarded as a model system<sup>3</sup> for quantum electrodynamics (for an introduction see, e.g., ref 12). Here, we demonstrate that even small molecules can serve to illustrate features of the Dirac equation subject to finite-system boundary conditions. There appears to be a considerable similarity between finite conjugated systems, such as the ones discussed here, and graphene. It is puzzling that involved ideas of quantum electrodynamics can be illustrated in terms of the Hückel model of polyenes. However, we should add that, in finite systems, there are boundary effects, such as edge states, discussed in the previous section, that have no analogue in graphene. Furthermore, graphene is a two-dimensional (2D) system, and therefore, there are topological effects that have no counterpart in one-dimensional (1D) systems. An example

is the Berry phase that the wave function of graphene can acquire if the wave vector varies along closed loops.<sup>4</sup>

The free-electron model is often employed to qualitatively explain the behavior of  $\pi$ -electrons in polyene. For example, this model has been used recently<sup>34</sup> to investigate the electron transport properties of conjugated systems. Since the free-electron model is based on a differential equation and not on a matrix equation, its solutions can be obtained analytically. However, its quadratic dispersion relation does not correctly reproduce the energy level spacing of polyenes around the highest occupied molecular orbital/lowest unoccupied molecular orbital (HOMO/LUMO) energies. The discrete Dirac equation presented here results in an alternative continuum model given by the 1D Dirac equation. This equation yields a linear dispersion relation and rectifies a problem of the free-electron model. Furthermore, the Dirac energy spectrum is symmetric with respect to the midpoint between the HOMO and the LUMO. This is a feature of the Hückel spectrum as well that is not reproduced by the conventional free-electron model.

**Acknowledgment.** We would like to thank P. Rocheleau whose critical reading of the manuscript lead to various improvements. M.E. is indebted to Didier Mayou for inspiring discussions which sparked his interest in graphene. Furthermore, we gratefully acknowledge financial support provided by Natural Sciences and Engineering Research Council (NSERC).

## References

- (1) Novoselov, K. S.; Geim, A. K.; Morozov, S. V.; Jiang, D.; Zhang, Y.; Dubonos, S. V.; Grigorieva, I. V.; Firsov, A. A. *Science* **2004**, *306*, 666–669.
- (2) Geim, A. K.; Novoselov, K. S. *Nat. Mater.* **2007**, *6*, 183–191.
- (3) Geim, A. K. *Science* **2009**, *324*, 1530–1534.
- (4) Neto, A. H. C.; Guinea, F.; Peres, N. M. R.; Novoselov, K. S.; Geim, A. K. *Rev. Mod. Phys.* **2009**, *81*, 109–162.
- (5) Rao, C.; Sood, A.; Subrahmanyam, K.; Govindaraj, A. *Angew. Chem., Int. Ed.* **2009**, *48*, 7752–7777.
- (6) Wallace, P. R. *Phys. Rev.* **1947**, *71*, 622–634.
- (7) Semenoff, G. W. *Phys. Rev. Lett.* **1984**, *53*, 2449–2452.
- (8) DiVincenzo, D. P.; Mele, E. J. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1984**, *29*, 1685–1694.
- (9) Novoselov, K. S.; Jiang, Z.; Zhang, Y.; Morozov, S. V.; Stormer, H. L.; Zeitler, U.; Maan, J. C.; Boebinger, G. S.; Kim, P.; Geim, A. K. *Science* **2007**, *315*, 1379–1379.
- (10) Ozyilmaz, B.; Jarillo-Herrero, P.; Efetov, D.; Abanin, D. A.; Levitov, L. S.; Kim, P. *Phys. Rev. Lett.* **2007**, *99*, 166804.
- (11) Bunch, J. S.; Yaish, Y.; Brink, M.; Bolotin, K.; McEuen, P. L. *Nano Lett.* **2005**, *5*, 287–290.
- (12) Ryder, L. H. *Quantum Field Theory*; Cambridge University Press: Cambridge, U.K., 1985.
- (13) Matulis, A.; Peeters, F. M. *Am. J. Phys.* **2009**, *77*, 595–601.
- (14) Coulson, C. A. *Proc. Phys. Soc., London, Sect. A* **1953**, *66*, 652–655.
- (15) Malysheva, L.; Onipko, A. *Phys. Rev. Lett.* **2008**, *100*, 186806.

- (16) Onipko, A. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2008**, *78*, 245412.
- (17) Nikolaev, A. V.; Bibikov, A. V.; Avdeenkova, A. V.; Bodrenko, I. V.; Tkalya, E. V. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2009**, *79*, 045418.
- (18) Hod, O.; Peralta, J. E.; Scuseria, G. E. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2007**, *76*, 233401.
- (19) Ezawa, M. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2007**, *76*, 245415.
- (20) Hod, O.; Barone, V.; Peralta, J. E.; Scuseria, G. E. *Nano Lett.* **2007**, *7*, 2295–2299.
- (21) Fernández-Rossier, J.; Palacios, J. J. *Phys. Rev. Lett.* **2007**, *99*, 177204.
- (22) Shemella, P.; Zhang, Y.; Mailman, M.; Ajayan, P. M.; Nayak, S. K. *Appl. Phys. Lett.* **2007**, *91*, 042101.
- (23) Ezawa, M. *Phys. E (Amsterdam, Neth.)* **2008**, *40*, 1421–1423. 17th International Conference on Electronic Properties of Two-Dimensional Systems.
- (24) Hod, O.; Scuseria, G. E. *ACS Nano* **2008**, *2*, 2243–2249.
- (25) Kudin, K. N. *ACS Nano* **2008**, *2*, 516–522.
- (26) Hod, O.; Barone, V.; Scuseria, G. E. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2008**, *77*, 035411.
- (27) Jiang, D.; Chen, X.-Q.; Luo, W.; Shelton, W. A. *Chem. Phys. Lett.* **2009**, *483*, 120–123.
- (28) Fujita, M.; Wakabayashi, K.; Nakada, K.; Kusakabe, K. *J. Phys. Soc. Jpn.* **1996**, *65*, 1920–1923.
- (29) Wakabayashi, K.; Fujita, M.; Ajiki, H.; Sigrist, M. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1999**, *59*, 8271–8282.
- (30) Son, Y.-W.; Cohen, M. L.; Louie, S. G. *Phys. Rev. Lett.* **2006**, *97*, 216803.
- (31) Zheng, H.; Wang, Z. F.; Luo, T.; Shi, Q. W.; Chen, J. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2007**, *75*, 165414.
- (32) Barone, V.; Hod, O.; Scuseria, G. E. *Nano Lett.* **2006**, *6*, 2748–2754.
- (33) Son, Y.-W.; Cohen, M. L.; Louie, S. G. *Phys. Rev. Lett.* **2006**, *97*, 216803.
- (34) Hsu, L.-Y.; Jin, B.-Y. *Chem. Phys.* **2009**, *355*, 177–182.

CT1000044



## Using Nonempirical Semilocal Density Functionals and Empirical Dispersion Corrections to Model Dative Bonding in Substituted Boranes

Benjamin G. Janesko\*

Department of Chemistry, Texas Christian University Fort Worth, Texas 76109

Received February 11, 2010

**Abstract:** Dative bonds to substituted boranes represent a challenge for the approximate exchange-correlation functionals typically used in density functional theory (DFT). Accurately modeling these bonds with DFT has usually required highly parametrized functionals, large admixtures of exact exchange, or computationally expensive double hybrids. This work shows that the nonempirical semilocal PBEsol functional, and the nonempirical semilocal PBE and TPSS functionals augmented with empirical interatomic dispersion corrections, accurately treat several representative problems in dative bonding. These methods typically surpass the MPW1K “kinetics” global hybrid previously recommended for dative bonds. This work also provides additional insights into the accuracy of the parametrized M06 functionals and indicates some deficiencies of the B97-D functional relative to PBE-D and TPSS-D. Applications to frustrated Lewis pairs illustrate the potential of these methods.

### 1. Introduction

Density functional theory (DFT) incorporating approximate exchange-correlation functionals has become a favored approach for modeling chemical bonding in condensed phases and medium-sized to large molecules.<sup>1</sup> But recent investigations<sup>2–4</sup> have demonstrated that conventional functionals have severe problems treating noncovalent interactions. This has stimulated several methodological developments in DFT. These include empirical<sup>5,6</sup> and nonempirical<sup>7,8</sup> interatomic dispersion corrections, effective atomic core potentials to model dispersion,<sup>9</sup> “fifth-rung” functionals<sup>10</sup> incorporating approximate second-order Görling-Levy perturbation theory correlation,<sup>11</sup> and semiempirical functionals containing a large number of fitted parameters.<sup>12</sup>

Dative bonds (coordinate covalent bonds) to substituted boranes form an important class of relatively weak interactions. Dative bonds are strongly influenced by factors such as crystal packing<sup>13–15</sup> and the interplay of substituents’ steric and electronic effects.<sup>16</sup> Boron’s dative bonds play important roles in areas including sensing<sup>17,18</sup> and supramolecular chemistry.<sup>19</sup> Recent reports of heterolytic H<sub>2</sub> splitting

by “frustrated Lewis pairs” between sterically hindered boranes and Lewis bases<sup>20,21</sup> has engendered much computational work on these systems.<sup>22–27</sup> More broadly, dative bonding is central to transition metal chemistry, another focus of recent computational work.<sup>28–32</sup>

Dative bonds to substituted boranes are a significant challenge for DFT. Standard DFT approximations, including the B3LYP<sup>33,34</sup> functional used in previous investigations of borane dative bonding,<sup>18,35</sup> often display qualitative and quantitative failures for these systems. Gilbert showed that B3LYP cannot reproduce accurate B–N bond lengths and bond dissociation energies for several Me<sub>n</sub>H<sub>3–n</sub>B–NMe<sub>m</sub>H<sub>3–m</sub> and (CF<sub>3</sub>)<sub>n</sub>H<sub>3–n</sub>B–NMe<sub>m</sub>H<sub>3–m</sub> species (ref 36, “Me” = CH<sub>3</sub>). Particularly egregious failures occurred in the sort of highly substituted, sterically congested molecules relevant to frustrated Lewis pairing. The best DFT results were obtained with the kinetics global hybrid MPW1K,<sup>37</sup> which is parametrized to reproduce reaction barrier heights. This led the author to conclude that methods designed for “incompletely bound” transition states are most appropriate for modeling dative bonds. Phillips and Cramer found that hybrid DFT functionals were required to model the bond length in F<sub>3</sub>B–NCH and that even hybrids could not reproduce accurate B–N bond energies.<sup>38</sup> Plumley and

\* To whom correspondence should be addressed; E-mail: b.janesko@tcu.edu.

Evansack found B3LYP to be completely inadequate for modeling substituent effects in  $\text{Me}_3\text{B}-\text{NMe}_n\text{H}_{3-n}$  complexes.<sup>39</sup> In these systems, competition between steric and electronic effects makes the B–N bond enthalpy increase for  $n = 0 \rightarrow 2$  and then decrease at  $n = 3$ .<sup>40</sup> Of the functionals tested by the authors, only the highly parametrized Minnesota functionals<sup>12</sup> reproduced this experimental trend. The authors recommended the M06-2X functional, which they used extensively in a subsequent study of boron's Lewis acidity.<sup>16</sup> Rakow and co-workers found that empirically dispersion-corrected functionals, or Minnesota functionals, were needed to treat H/Br exchange barriers in  $\text{BBr}_3$ .<sup>41</sup>

This state of affairs is somewhat unsatisfying. Global hybrid density functionals like MPW1K and M06-2X have a relatively large computational cost due to their inclusion of exact (Hartree–Fock-type, HF) exchange.<sup>1</sup> Indeed, a recent MPW1K treatment of the frustrated Lewis pair and  $\text{H}_2$  activating complex<sup>20</sup>  $(\text{C}_6\text{F}_5)_3\text{B}-\text{P}(\text{tBu})_3$  [“tBu” =  $\text{C}(\text{CH}_3)_3$ ] used the ONIOM embedding method<sup>42</sup> due to computational cost.<sup>23</sup> The long-range part of hybrids' HF exchange is especially problematic in condensed phases,<sup>43</sup> making global hybrids inappropriate for dative bonding in nanostructures and condensed phases. Fifth-rung functionals like the B2PLYP-D<sup>44</sup> method recommended in ref 41 have an even higher computational expense. Additionally, while the M06 suite of functionals has demonstrated broad utility in chemistry,<sup>12</sup> its members have a large number of empirical parameters and appear prone to numerical errors.<sup>45,46</sup>

There would be great value in a semilocal density functional that could accurately treat dative bonds in molecules, nanostructures, and condensed phases, with a low computational cost and a minimum of empiricism. (“Semilocal” density functionals model the exchange-correlation energy density and potential at each point  $\mathbf{r}$  as a function of the electron density, density gradient, and possibly the noninteracting kinetic energy density and/or density Laplacian at  $\mathbf{r}$ .<sup>1</sup> They are typically computationally cheaper than global hybrids.) Here, I explore two recent approximations with these characteristics: the nonempirical PBEsol generalized gradient approximation (GGA)<sup>47</sup> and the addition of empirical dispersion corrections<sup>5</sup> to nonempirical semilocal functionals.

Surprising recent investigations have found that simple, nonempirical density functionals can accurately model the “medium-range correlation” important to noncovalent interactions. Woodrich and co-workers showed that the local spin-density approximation (LSDA) outperformed conventional functionals for isodesmic stabilization energies of  $n$ -alkanes.<sup>3</sup> Csonka and co-workers showed that PBEsol,<sup>47</sup> while designed for solids, accurately models a range of noncovalent stereoelectronic effects.<sup>48</sup> B3LYP gives qualitative and quantitative failures for these properties.<sup>2</sup> PBEsol has the form of the Perdew–Burke–Ernzerhof (PBE) GGA<sup>49</sup> and modifies two parameters to restore the correct density gradient expansion for exchange.<sup>47</sup>

Another important development is the advent of empirical interatomic dispersion corrections<sup>50</sup> in DFT.<sup>5,6,51</sup> Grimme's “-D” corrections add a damped, molecular-mechanics-type

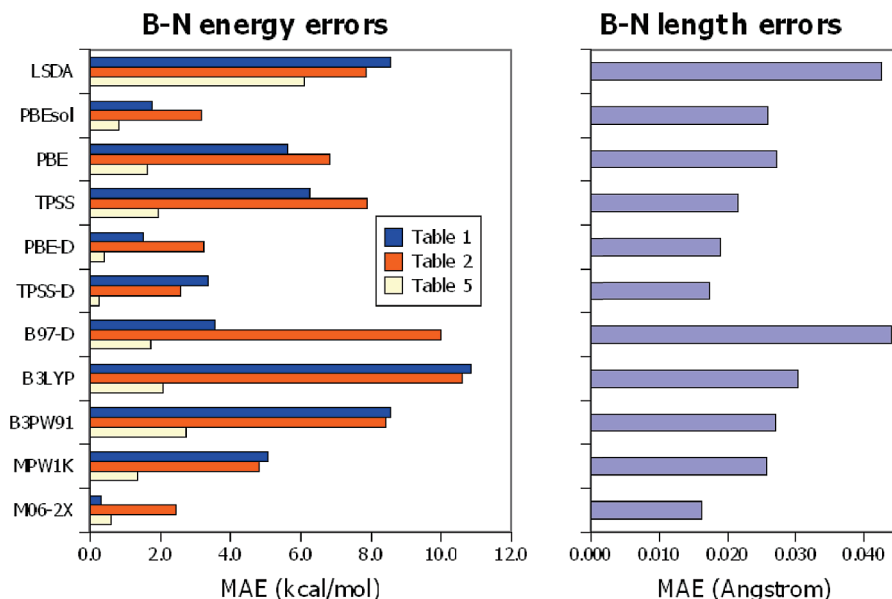
$R^{-6}$  internuclear attraction to a DFT calculation. These provide a straightforward route to improving noncovalent interactions in large systems.<sup>51</sup> They have been successfully applied to a few problems in dative bonding.<sup>22,26,41,52,53</sup> However, much of that work focused on dispersion-corrected double hybrids whose computational expense is comparable to MP2. To date, a systematic study of “-D” corrections for dative bonding has not appeared.

This work benchmarks the methods of refs 3, 48, and 51 for dative bonds to substituted boranes. I test the nonempirical LSDA and PBEsol functionals, and the addition of empirical dispersion corrections to the nonempirical PBE and TPSS functionals. These methods significantly improve upon conventional global hybrids, approaching the accuracy of M06-2X at reduced computational cost. Selected results for the entire M06 suite of functionals provide new insight into these methods' success. Conversely, tests of the dispersion-corrected B97-D functional indicate that its accurate performance in other areas<sup>51</sup> does not carry over to dative bonds.

## 2. Computational Details

All calculations use a development version of the Gaussian suite of programs.<sup>54</sup> Unless noted otherwise, calculations use the correlation-consistent aug-cc-pVTZ basis set<sup>55</sup> or the large Pople basis 6-311++G(3df,2p).<sup>56</sup> Binding energies and enthalpies include counterpoise corrections for basis set superposition error.<sup>57</sup> Self-consistent field (SCF) calculations converge the energy to at least  $10^{-8}$  Hartree with corresponding thresholds for geometry calculations (Gaussian keywords “SCF=Tight” and “Geom=Tight”). DFT numerical integrations use at least an “UltraFine” integration grid with 99 radial and 590 angular points per atom. Unless noted otherwise, energies are evaluated with geometries optimized using the corresponding method. Counterpoise corrections are not included during the geometry optimizations. All calculations treat isolated molecules, with no corrections for solvation or crystal packing effects. Open-shell systems are treated spin-unrestricted. Other computational details are taken from the literature (vide infra).

This work focuses on nonempirical and dispersion-corrected semilocal exchange-correlation functionals. These include the local spin-density approximation LSDA (Vosko–Wilk–Nusair correlation functional V, ref 58), the Perdew–Burke–Ernzerhof (PBE) GGA,<sup>49</sup> the Tao–Perdew–Staroverov–Scuseria (TPSS) meta-GGA,<sup>59</sup> and the PBEsol GGA.<sup>47</sup> PBE-D and TPSS-D use Grimme's empirical interatomic dispersion correction.<sup>5,6,51</sup> Scale factors  $s_6 = 0.75$  (PBE-D) and  $s_6 = 1.00$  (TPSS-D) are taken from ref 6. The empirical B97-D functional is also tested.<sup>6</sup> Selected systems are tested for the entire M06 suite of functionals: the semilocal M06-L meta-GGA,<sup>60</sup> the M06 hybrid meta-GGA incorporating 27% HF exchange,<sup>12</sup> the M06-2X hybrid with 54% HF exchange,<sup>12</sup> and the “density functional for spectroscopy” M06-HF with 100% HF exchange.<sup>61</sup> While M06-2X is recommended for problems like dative bonding,<sup>12</sup> results from the entire suite can provide interesting insights into these functionals. The B3PW91,<sup>33</sup> B3LYP,<sup>34</sup> and MPW1K<sup>37</sup> global hybrid functionals are included for comparisons to previous work.



**Figure 1.** Cumulative statistical errors in B–N bond strengths (left) and bond lengths (right). The left panel plots the MAE in B–N bond energies/enthalpies from Tables 1, 2, and 5. The right panel plots the MAE in B–N bond lengths from Table 4.

**Table 1.** B–N Bond Enthalpies at 373 K (kcal/mol) for Methyl Substituted Amine Boranes<sup>a</sup>

method	Me <sub>3</sub> B–NH <sub>3</sub>	Me <sub>3</sub> B–NH <sub>2</sub> Me	Me <sub>3</sub> B–NHMe <sub>2</sub>	Me <sub>3</sub> B–NMe <sub>3</sub>	MAE
exptl	–13.8 ± 0.3	–17.6 ± 0.2	–19.3 ± 0.3	–17.6 ± 0.2	
LSDA	–22.6	–27.1	–27.6	–25.2	8.6
PBEsol	–16.6	–19.7	–19.1	–15.7	1.8
PBE	–11.4	–13.9	–12.4	–8.1	5.6
TPSS	–11.0	–13.0	–11.6	–7.7	6.3
PBE-D	–15.2	–19.7	–20.7	–18.8	1.5
TPSS-D	–16.0	–20.9	–22.7	–22.2	3.4
B97-D	–9.8	–14.4	–15.6	–14.3	3.6
B3LYP	–6.8	–8.8	–7.0	–2.4	10.8
B3PW91	–9.1	–11.0	–9.2	–4.8	8.6
MPW1K	–11.9	–14.2	–12.9	–9.1	5.1
M06-L	–9.7	–13.1	–13.3	–11.5	5.2
M06	–10.3	–14.0	–14.6	–12.5	4.2
M06-2X	–14.1	–18.3	–19.2	–17.8	0.3
M06-HF	–20.0	–24.9	–26.7	–26.2	7.4

<sup>a</sup> Counterpoise-corrected 6-311++G(3df,2p) calculations, experimental results from ref 39.

### 3. Results

Figure 1 presents a cumulative picture of the results. The figure shows mean absolute errors (MAE) in bond energies/enthalpies and bond lengths for several sets of B–N dative bonds. To summarize, the dispersion-corrected PBE-D and TPSS-D functionals give very accurate B–N dissociation energies and bond lengths. These methods surpass the previously recommended MPW1K functional and approach the accurate M06-2X functional. PBEsol is also accurate for B–N dissociation energies, though it tends to underestimate bond lengths. The remainder of this section details the individual studies in Figure 1 and discusses applications to larger systems.

#### 3.1. Bond Strengths of Substituted Amine Boranes.

References 39 and 40 studied the B–N bond enthalpies of four Me<sub>3</sub>B–NMe<sub>n</sub>H<sub>3–n</sub> derivatives. These references tested the B3LYP, MPW1K, MPWB1K, MPW1B95, and M05/M06 density functionals. The experimental trend, in which the B–N dative bond strengthens for *n* from 0–2 and

decreases at *n* = 3, was only reproduced by the M05-2X, M06, and M06-2X functionals.

Table 1 shows new calculations on these Me<sub>3</sub>B–NMe<sub>n</sub>H<sub>3–n</sub> derivatives. The calculations in Table 1 follow refs 39 and 40, evaluating binding enthalpies at 373 K using B3LYP/6-31G(d) thermal corrections rescaled by 0.9941. Counterpoise-corrected thermal corrections were taken from ref 39. The B3LYP, MPW1K, and M06-2X results in Table 1 reproduce ref 40 to within ±0.1 kcal/mol. The basis set is relatively saturated: all methods except for the Minnesota functionals give counterpoise corrections < 1 mH.

The most striking results in Table 1 involve the relative binding enthalpies. While LSDA overbinds, it correctly predicts that the bond enthalpies strengthen for *n* from 0 to 2 and decrease at *n* = 3. None of the DFT methods tested in ref 39 reproduced this trend. PBEsol nearly reproduces the trend and also gives absolute bond enthalpies that significantly improve on LSDA. This is consistent with previous observations of medium-range correlation in LSDA<sup>3</sup> and

**Table 2.** B–N Bond Dissociation Energies (kcal/mol) for Trifluoromethyl Substituted Amine Boranes<sup>a</sup>

method	(CF <sub>3</sub> )H <sub>2</sub> B–NH <sub>3</sub>	(CF <sub>3</sub> ) <sub>2</sub> HB–NH <sub>3</sub>	(CF <sub>3</sub> ) <sub>3</sub> B–NH <sub>3</sub>	MAE
reference	–40.4	–52.6	–62.6	
LSDA	–51.7	–60.1	–67.4	7.9
PBEsol	–44.3	–51.8	–57.8	3.2
PBE	–38.3	–45.7	–51.1	6.8
TPSS	–36.7	–44.6	–50.6	7.9
PBE-D	–41.0	–49.7	–56.4	3.2
TPSS-D	–40.3	–50.0	–57.6	2.6
B97-D	–33.6	–42.6	–49.4	10.0
B3LYP	–33.5	–42.0	–48.3	10.6
B3PW91	–36.1	–44.1	–50.1	8.4
MPW1K	–38.6	–47.8	–54.8	4.8
M06-2X	–39.2	–49.9	–59.1	2.5

<sup>a</sup> Counterpoise-corrected 6-311++G(3df,2p) calculations. Reference MP2/6-311++G(d,p) results are from ref 36.

PBEsol.<sup>48</sup> These results are not a basis set artifact: PBEsol calculations in the aug-cc-pVTZ basis set return bond enthalpies within  $\pm 0.1$  kcal/mol of those in Table 1. PBEsol significantly improves upon the MPW1K functional recommended in ref 36 and approaches the accuracy of M06-2X.

Another important result in Table 1 is the success of empirical dispersion corrections. PBE and TPSS do not reproduce the absolute bond enthalpies of Me<sub>3</sub>B–NMe<sub>n</sub>H<sub>3–n</sub> or the trend with increasing *n*. But adding empirical “-D” dispersion corrections to these nonempirical functionals dramatically improves their performance. PBE-D and TPSS-D both reproduce the experimental trend, and PBE-D gives overall accuracy comparable to PBEsol and approaching M06-2X. This result agrees with ref 62, which showed that PBE-D reproduces MP2 benchmarks for H<sub>3</sub>B–PH<sub>3</sub> and Me<sub>3</sub>B–PMe<sub>3</sub>.

The B97-D GGA, whose empirical functional form was explicitly parametrized to complement its dispersion correction,<sup>6</sup> underestimates B–N bond strengths. This is unlikely to be a basis set artifact: geometry-optimized B97-D/aug-cc-pVTZ calculations give counterpoise-corrected dissociation enthalpies within  $\pm 0.1$  kcal/mol of those in Table 1. B97-D also significantly overestimates B–N dative bond lengths (vide infra). Note in this context that ref 22 showed B97-D significantly underestimates the B–P dative bond strength in a cyclic intramolecular phosphane–borane adduct.<sup>63</sup>

Table 1 also shows that the entire M06 series of functionals reproduces the experimental trend in binding enthalpies. While the absolute binding enthalpies vary with the fraction of HF exchange, all four functionals appear to give a reasonable account of the medium-range correlation that produces this trend. This indicates that the parametrization procedure for these functionals is quite robust.

Table 2 shows the B–N bond dissociation energies of trifluoromethyl substituted amine boranes (CF<sub>3</sub>)<sub>n</sub>H<sub>3–n</sub>B–NH<sub>3</sub>. These compounds were studied in ref 36, which concluded that conventional DFT functionals failed to reproduce the steep increase in bond energy upon fluorination. The best DFT results were obtained with the MPW1K kinetics global hybrid. This was justified by the observation that MPW1K

is designed to model transition states, and by the claim that “incompletely bound” transition states mimic datively bonded systems.<sup>36</sup>

Calculations in Table 2 use the 6-311++G(3df,2p) basis. They include HF/6-31+G(d) zero-point energy corrections empirically rescaled by 0.9153, following ref 36. The counterpoise corrections are somewhat larger than in Table 1, though they are  $< 2$  mH in all but a few systems. The B3LYP, B3PW91, and MPW1K bond energies are a few kilocalories per mole smaller than the corresponding non-counterpoise-corrected 6-311++G(d,p) values in ref 36. Aug-cc-pVTZ calculations give counterpoise-corrected PBE-D, B97-D, and M06-2X dissociation energies within  $\pm 0.2$  kcal/mol of Table 2. This indicates that the results are unlikely to be a basis set artifact.

The most important result in Table 2 is the accuracy of the PBEsol GGA and the dispersion-corrected PBE-D and TPSS-D functionals. As in Table 1, these methods all improve upon MPW1K at this level of theory and approach M06-2X. Table 2 also supports refs 40 and 41 in demonstrating that M06-2X is very accurate for these systems. As in Table 1, B97-D significantly underbinds relative to PBE-D and TPSS-D.

The other nonempirical functionals in Table 2 perform less well. LSDA strongly overbinds, as expected, and PBE and TPSS tend to underbind. But even these functionals, like BPW91, outperform B3LYP for this system.

The non-counterpoise-corrected MPW1K/6-311++G(d,p) results reported in ref 36 are better than any functionals in Table 2 in reproducing the non-counterpoise-corrected MP2/6-311++G(d,p) reference values. This is partly an artifact of a cancellation between finite basis set error and basis set superposition error (BSSE). Omitting the counterpoise correction gives PBEsol, PBE-D, TPSS-D, and M06-2X MAE of 2.7, 2.8, 2.0, and 1.7 kcal/mol vs the reference values in Table 2. These are comparable to the 1.9 kcal/mol MAE reported for MPW1K in ref 36. Additionally, new counterpoise-corrected MP2/aug-cc-pVTZ calculations at MP2/cc-pVTZ geometries give binding energies of –37.9, –48.6, and –57.3 kcal/mol for the three molecules in Table 2. These are 3–5 kcal/mol below the corresponding noncounterpoise-corrected MP2/6-311++G(2d,2p) values in ref 36. Comparing the calculations in Table 2 to these new reference values gives PBEsol, PBE-D, TPSS-D, MPW1K, and M06-2X MAE of 3.4, 1.7, 1.4, 1.3, and 1.4 kcal/mol. PBE-D, TPSS-D, and M06-2X are again comparable to MPW1K.

The accurate performance of PBEsol, PBE-D, and TPSS-D in Tables 1 and 2 suggests that accurate performance for “incompletely bound” transition states is not necessary for modeling dative bonds. Unlike MPW1K and M06-2X, these functionals contain no HF exchange and cannot adequately predict gas-phase reaction barriers. This is illustrated in Table 3, which shows statistical errors in the BH6 set of representative hydrogen-transfer reaction barrier heights.<sup>64</sup> Geometries for this data set are from ref 64, and spin–orbit corrections are from ref 65. PBEsol’s poor performance for reaction barriers was also shown in ref 66. I suggest that a method’s description of medium-range correlation, and not



**Table 3.** Mean Errors ME and Mean Absolute Errors MAE (kcal/mol) in BH<sub>6</sub> Hydrogen-Transfer Reaction Barrier Heights<sup>a</sup>

method	ME	MAE
LSDA	-17.9	17.9
PBEsol	-12.8	12.8
PBE	-9.4	9.4
PBE-D	-9.8	9.8
TPSS-D	-8.8	8.8
B97-D	-6.0	6.3
MPW1K	-1.1	1.4
M06-2X	-0.7	1.2

<sup>a</sup> Aug-cc-pVTZ calculations.**Table 4.** Equilibrium B–N Bond Lengths (Å) for Methyl Substituted Amine Boranes<sup>a</sup>

method	H <sub>3</sub> B–NH <sub>3</sub>	H <sub>3</sub> B–NMe <sub>3</sub>	Me <sub>3</sub> B–NMe <sub>3</sub>	MAE
exptl	1.658(2)	1.656(2)	1.70(1)	
LSDA	1.606	1.599	1.682	0.043
PBEsol	1.628	1.621	1.714	0.026
PBE	1.647	1.642	1.758	0.027
TPSS	1.663	1.653	1.757	0.021
PBE-D	1.653	1.637	1.734	0.019
TPSS-D	1.672	1.645	1.728	0.017
B97-D	1.692	1.665	1.790	0.044
B3LYP	1.657	1.651	1.785	0.030
B3PW91	1.644	1.637	1.749	0.027
MPW1K	1.632	1.625	1.722	0.026
M06-2X	1.648	1.638	1.722	0.016

<sup>a</sup> 6-311++G(3df,2p) calculations. Gas-phase experimental values are taken from ref 36.

its treatment of transition states, is the decisive factor for modeling dative bonding.

**3.2. Geometries of Substituted Amine Boranes.** The B–N bond length in amine boranes and related compounds<sup>67</sup> is sensitive to crystal packing effects,<sup>13–15</sup> making it inappropriate to compare gas-phase calculations with experimental crystal structures. Reference 36 collected gas-phase experimental structures for three substituted amine boranes. That reference concluded that, as for bond energies, MPW1K provided the most accurate treatment of bond lengths.

Table 4 shows how several DFT methods reproduce these experiments. The table shows the equilibrium B–N bond lengths from 6-311++G(3df,2p) calculations on H<sub>3</sub>B–NH<sub>3</sub>, H<sub>3</sub>B–NMe<sub>3</sub>, and Me<sub>3</sub>B–NMe<sub>3</sub>. The B3LYP and MPW1K results are within 0.01 Å of the 6-311++G(d,p) results in ref 36. Test calculations indicate that the results are unlikely to be a basis set artifact. Aug-cc-pVTZ calculations with PBEsol, PBE-D, B97-D, and M06-2X give bond lengths 0.002 Å longer than those in Table 4 for H<sub>3</sub>B–NH<sub>3</sub>, and within ±0.001 Å of those in Table 4 for the other two molecules.

The accurate performance of PBE-D, TPSS-D, and M06-2X for dative bond energies carries over to bond lengths. These methods provide the most accurate results in Table 4. All three outperform both the MPW1K/6-311++G(3df,2p) geometries in Table 4 and the MPW1K/6-311++G(d,p) geometries reported in ref 36. While TPSS gives the lowest MAE among nonempirical functionals, it overestimates the bond length of the most highly substituted compound.

**Table 5.** Equilibrium B–N Bond Lengths (*R*(B–N), Å) and Counterpoise-Corrected B–N Dissociation Energies (DE, kcal/mol) for HCN–BF<sub>3</sub><sup>a</sup>

method	<i>R</i> (B–N)	DE
reference	2.472	-5.6
LSDA	1.729	-11.7
PBEsol	1.820	-6.4
PBE	2.410	-4.0
TPSS	2.243	-3.7
PBE-D	2.362	-5.2
TPSS-D	2.234	-5.3
B97-D	2.682	-3.9
B3LYP	2.535	-3.5
B3PW91	2.465	-2.9
MPW1K	2.323	-4.2
M06-2X	2.385	-6.2

<sup>a</sup> Aug-cc-pVTZ calculations without zero-point or thermal corrections. Reference MC-QCISD results are from ref 38.

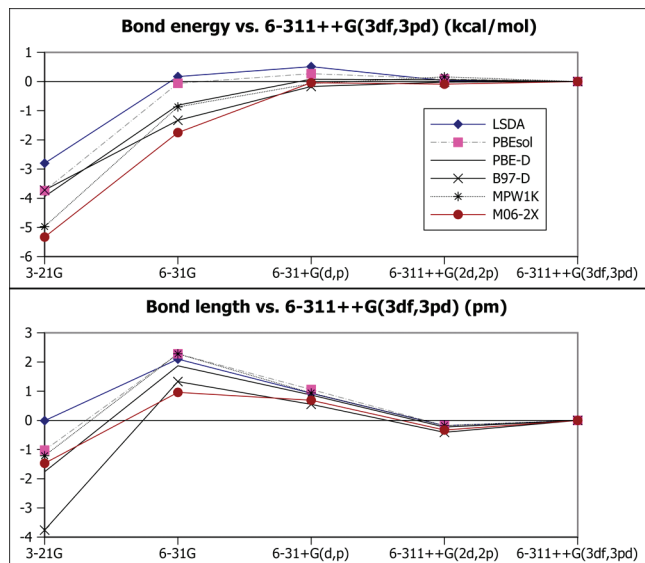
(Similar errors for conventional DFT functionals were shown in ref 36.) This overestimation is improved by the “-D” dispersion correction, illustrating the importance of medium-range correlation in this highly substituted system.

As mentioned above, B97-D overestimates all of the B–N bond lengths, giving the largest MAE of any tested functional. The accurate PBE-D and TPSS-D results suggest that this is not a failure of the dispersion correction but a limitation of the B97-D parametrization. This result rationalizes B97-D’s poor performance for the bond energies/enthalpies in Tables 1 and 2. (Here, it is appropriate to reiterate B97-D’s accuracy for other systems.<sup>51</sup>)

**3.3. Bond Strength and Bond Length in HCN–BF<sub>3</sub>.** Reference 38 characterized the gas-phase structure and B–N binding energies of the HCN–BF<sub>3</sub> dative bond. Comparisons to accurate MC-QCISD or MG3 results showed that, while some conventional functionals gave reasonable geometries, all tended to underestimate bond energies. Table 5 shows HCN–BF<sub>3</sub> B–N bond lengths and dissociation energies for the functionals considered here. These counterpoise-corrected aug-cc-pVTZ results differ slightly from the non-counterpoise-corrected results in ref 38. The counterpoise corrections are <1 mH, indicating that the basis set is reasonably saturated.

The results in Table 5 reiterate those in Tables 1–4. PBE-D, TPSS-D, and M06-2X accurately model both the bond length and the dissociation energy, providing significant improvements over the B3PW91, B3LYP, and MPW1K global hybrids. PBEsol strikes a balance between the overbinding of LSDA and the underbinding of PBE, though (as in Table 4), it underestimates the dative bond length. Unlike the other dispersion-corrected functionals, B97-D severely overestimates the bond length and underestimates the bond energy. The spread in bond length errors is much larger than in Table 4, indicating that the bond has a shallow minimum.

**3.4. Basis Set Dependence.** Previous studies have indicated that dative bonds strongly depend on the one-electron basis set.<sup>36,38</sup> Given this, it is of interest to test the basis set dependence of the density functionals used here. Figure 2 shows the basis set dependence of a representative system: the equilibrium B–N bond length and counterpoise-corrected



**Figure 2.** Basis set dependence of B–N bond energy (top) and bond length (bottom) in  $(\text{CH}_3)_3\text{B}-\text{N}(\text{CH}_3)_3$ . Counterpoise-corrected geometry-optimized calculations. Results are relative to the large 6-311++G(3df,3pd) basis set.

bond energy of  $\text{Me}_3\text{B}-\text{NMe}_3$ . The figure shows the bond energy (top) and bond length (bottom) evaluated in a variety of basis sets. Results are evaluated relative to the large 6-311++G(3df,3pd) basis set. All of the methods tested here have roughly comparable basis set dependence. This is encouraging given previous reports<sup>41</sup> of large basis set effects for the Minnesota functionals. All methods show a rather large basis set dependence for bond lengths. However, bond lengths and bond energies are fairly well converged in the “desert island double- $\zeta$ ”<sup>68</sup> 6-31+G(d,p) basis set. This basis set should provide a reasonable compromise for modeling larger systems such as frustrated Lewis pairs.

**3.5. Application to Larger Systems.** One goal of this work is to find computationally inexpensive DFT methods for modeling the large, sterically congested substituted boranes relevant to frustrated Lewis pairing.<sup>21</sup> An important drawback of conventional semilocal and hybrid density functionals is that their errors tend to increase with system size. For example, the mean absolute error in B3LYP enthalpies of formation increases from 3.08 kcal/mol for the G2/97 test set of 147 small molecules, to 8.21 kcal/mol for the 75 larger molecules in the G3-3 set.<sup>69</sup> However, the parametrized and dispersion-corrected functionals tested here were constructed to provide comparable accuracy for both small and large systems. These functionals’ accuracy for large systems has been amply demonstrated in the literature. Reference 48 showed that M05-2X, PBEsol, and empirically dispersion-corrected functionals accurately predicted isomerization energies of large organic molecules. Comparable performance for M06-2X was demonstrated in ref 70. Reference 71 showed that dispersion-corrected semilocal functionals accurately predict intermolecular interaction energies for a data set containing both small (e.g.,  $(\text{H}_2\text{O})_2$ ) and relatively large (e.g., phenylalanine-tryptophan) biologically relevant complexes.<sup>4</sup> (Further applications of empirical dispersion corrections to large molecules are reviewed in ref

**Table 6.** Equilibrium B–P Bond Lengths ( $R(\text{B}-\text{P})$ , Å) and Counterpoise-Corrected B–P Bond Energies (DE, kcal/mol) for  $(\text{F}_5\text{C}_6)_3\text{B}-\text{P}(t\text{-Bu})_3$  and  $(\text{F}_3\text{C})_3\text{B}-\text{P}(t\text{-Bu})_3$ <sup>a</sup>

method	$(\text{F}_3\text{C})_3\text{B}-\text{P}(t\text{-Bu})_3$		$(\text{F}_5\text{C}_6)_3\text{B}-\text{P}(t\text{-Bu})_3$	
	DE	$R(\text{B}-\text{P})$	DE	$R(\text{B}-\text{P})$
reference	–69	2.163	–19	3.838
LSDA	–49	2.118	–12	3.476
PBEsol	–31	2.178	–1	4.123
PBE	–27	2.184	0	4.700
PBE-D	–45	2.133	–11	3.809
B97-D	–38	2.168	–10	3.963
MPW1K	–30	2.143	1	4.975
M06-2X	–44	2.122	–10	3.722

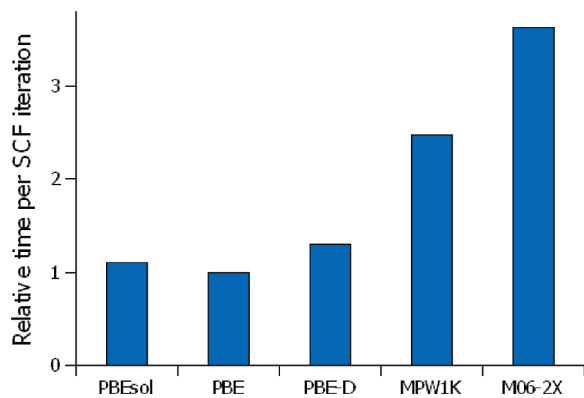
<sup>a</sup> 6-31+G(d,p) calculations, reference values from ref 23.

51.) Reference 70 showed that M06-2X gives a mean unsigned error of only 2.86 kcal/mol for the aforementioned G3-3 set of larger molecule thermochemistries, as well as a mean unsigned error of 5.7 kcal/mol (vs 26.2 kcal/mol for B3LYP) atomization energies. Given this, it seems reasonable to expect that the accurate performance shown above for M06-2X and PBE-D will carry over to larger molecules.

Table 6 illustrates some of the methods tested here for two relatively large systems.  $(\text{F}_5\text{C}_6)_3\text{B}-\text{P}(t\text{-Bu})_3$  ( $t\text{-Bu} = \text{C}(\text{CH}_3)_3$ ) is a weakly bound frustrated Lewis pair (ref 21) that performs heterolytic  $\text{H}_2$  splitting under relatively mild conditions.<sup>20</sup>  $(\text{F}_3\text{C})_3\text{B}-\text{P}(t\text{-Bu})_3$  was predicted in ref 23 to have a B–P dissociation energy of 69 kcal/mol, which is very high for a “weak” dative bond. These systems were modeled in ref 23 using the composite three-layer ONIOM G2R3 method at two-layer ONIOM MPW1K geometries.<sup>72,73</sup>

Table 6 presents 6-31+G(d,p) calculations of the equilibrium B–P bond length and counterpoise-corrected bond energy of  $(\text{F}_5\text{C}_6)_3\text{B}-\text{P}(t\text{-Bu})_3$  and  $(\text{F}_3\text{C})_3\text{B}-\text{P}(t\text{-Bu})_3$ . Calculations include HF/3-21G zero-point corrections rescaled by 0.9207, following ref 23. The maximum counterpoise corrections are 2.8 kcal/mol for  $(\text{F}_3\text{C})_3\text{B}-\text{P}(t\text{-Bu})_3$  and 2.3 kcal/mol for  $(\text{F}_5\text{C}_6)_3\text{B}-\text{P}(t\text{-Bu})_3$ , indicating that the basis set is moderately saturated. PBE-D/6-311++G(2d,2p) calculations at the PBE-D/6-31+G(d,p) geometries give counterpoise-corrected binding energies within  $\pm 0.1$  kcal/mol of those in Table 6, providing further confidence in the results.

Interestingly, M06-2X and PBE-D provide similar geometric and energetic predictions for both systems in Table 6. The bond energies are significantly smaller than the high-level composite values reported in ref 23. This appears to arise from basis set superposition error in the composite method. Reference 23 reported BSSEs of 18.2 and 10.6 kcal/mol in the composite method’s dative bond dissociation energies for  $(\text{F}_3\text{C})_3\text{B}-\text{PPh}_3$  and  $(\text{F}_5\text{C}_6)_3\text{Al}-\text{P}(\text{CH}_3)_3$  and estimated an average 14–15 kcal/mol BSSE for all molecules tested. Simply transferring the  $(\text{F}_3\text{C})_3\text{B}-\text{PPh}_3$  BSSE to  $(\text{F}_3\text{C})_3\text{B}-\text{P}(t\text{-Bu})_3$  and the  $(\text{F}_5\text{C}_6)_3\text{Al}-\text{P}(\text{CH}_3)_3$  BSSE to  $(\text{F}_5\text{C}_6)_3\text{B}-\text{P}(t\text{-Bu})_3$  gives dissociation energies within  $\sim 6$  and  $\sim 3$  kcal/mol of the respective PBE-D values. Additional insight into BSSE may be obtained from the B–N dissociation energy of  $(\text{CF}_3)_3\text{B}-\text{N}(\text{CH}_3)_3$ , a somewhat smaller molecule treated in ref 23. That reference reported dissociation energies of 67 kcal/mol for noncounterpoise-corrected



**Figure 3.** Average time per SCF iteration for a single self-consistent 6-31+G(d,p) total energy calculation on  $(F_5C_6)_3B-P(t-Bu)_3$ . Timings are reported relative to PBE.

MP2/6-311++G(d,p) and 64 kcal/mol for the composite method. New MP2/6-311++G(d,p) calculations give a non-counterpoise-corrected dissociation energy of 66.47 kcal/mol, and a corresponding counterpoise-corrected dissociation energy of only 52 kcal/mol. Counterpoise-corrected MPW1K, PBE-D, and M06-2X calculations give dissociation energies of 49, 58, and 63 kcal/mol, respectively. PBE-D and M06-2X binding energies are somewhat larger than MP2, which is reasonable given that MP2 should underbind in this relatively small basis set.

As in Tables 1–5, PBE and MPW1K both predict dative bonds that are significantly weaker than M06-2X or PBE-D. The relatively weak PBEsol bond in  $(F_5C_6)_3B-P(t-Bu)_3$  is somewhat surprising and suggests that medium-range correlation plays an unusually large role. The 4.975 Å MPW1K bond length in  $(F_5C_6)_3B-P(t-Bu)_3$  is significantly longer than the 3.838 Å MPW1K/ONIOM bond length obtained in ref 23. This difference may result from the small (3-21G) ligand basis set used in ref 23. MPW1K calculations in the 6-31G(d) and 3-21G basis sets yield B–P bond lengths of 4.388 and 3.631 Å, respectively, for this system.

The  $(F_5C_6)_3B-P(t-Bu)_3$  molecule shown in Table 6 also provides an opportunity to illustrate the relative computational expense of these methods. Figure 3 shows the average time per SCF cycle in a single-point 6-31+G(d,p) calculation on this system.<sup>74</sup> Of course, computational times strongly depend on details of the implementation and hardware, and these results should be taken as no more than a rough guide. But this reiterates that hybrids like MPW1K and M06-2X typically have a computational cost significantly higher than dispersion-corrected semilocal functionals.

#### 4. Conclusions

Previous DFT studies of dative bonds to substituted boranes indicated systematic failures of standard approximate exchange-correlation functionals. The results presented here show that the nonempirical PBEsol GGA gives accurate energies and reasonable (though overbound) geometries for a range of dative bonds. This extends previous indications<sup>48</sup> that PBEsol, which was built to model condensed phases, can mimic chemically important “medium-range” electron correlation effects. Adding empirical dispersion corrections to the nonempirical PBE and TPSS functionals gives even

higher overall accuracy, while maintaining a modest computational cost. These contrast with the dispersion-corrected B97-D GGA, which significantly overestimates B–N bond lengths. These results also support and extend previous indications<sup>40,41</sup> that the Minnesota functionals are very accurate for these systems. The entire suite of M06 functionals shows particularly notable accuracy for relative binding trends. However, applications to larger systems reiterate the value of the computationally cheap “-D” methods.

These results should extend the options available for modeling dative bonds in large systems. Dispersion-corrected nonempirical semilocal functionals offer the best overall balance between cost and accuracy, with only a modest degree of empiricism. M06-2X provides the highest numerical accuracy in systems where global hybrids are affordable.

**Acknowledgment.** This work was supported by a startup grant from Texas Christian University.

#### References

- (1) Scuseria, G. E.; Staroverov, V. N. Progress in the development of exchange-correlation functionals. In *Theory and Applications of Computational Chemistry: The First 40 Years*; Dykstra, C. E., Frenking, G., Kim, K. S., Scuseria, G. E., Eds.; Elsevier: Amsterdam, 2005; pp 669–724.
- (2) Grimme, S. *Angew. Chem., Int. Ed.* **2006**, *45*, 4460.
- (3) Woodrich, M. D.; Corminbeuf, C.; Schleyer, P. v. R. *Org. Lett.* **2006**, *8*, 3631.
- (4) Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985.
- (5) Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463.
- (6) Grimme, S. *J. Comput. Chem.* **2006**, *27*, 1787.
- (7) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2005**, *122*, 154104.
- (8) Tkatchenko, A.; Scheffler, M. *Phys. Rev. Lett.* **2009**, *102*, 073005.
- (9) von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. *Phys. Rev. Lett.* **2004**, *93*, 153004.
- (10) Perdew, J. P.; Schmidt, K. Jacob’s Ladder of Density Functional Approximations for the Exchange-Correlation Energy. In *Density Functional Theory and its Application to Materials*; Van Doren, V., Van Alsenoy, C., Geerlings, P., Eds.; American Institute of Physics, 2001; pp 1–20.
- (11) Grimme, S. *J. Chem. Phys.* **2006**, *124*, 034108.
- (12) Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008**, *120*, 215.
- (13) Thorne, L. R.; Suenram, R. D.; Lovas, F. J. *J. Chem. Phys.* **1983**, *78*, 167.
- (14) Bühl, M.; Steinke, T.; von Ragu’e Schleyer, Boese, R. *Angew. Chem., Int. Ed.* **1991**, *30*, 1160.
- (15) Finze, M.; Bernhardt, E.; Terheiden, A.; Berkei, M.; Willner, H.; Christen, D.; Oberhammer, H.; Aubke, F. *J. Am. Chem. Soc.* **2002**, *124*, 15385.
- (16) Plumley, J. A.; Evanseck, J. D. *J. Phys. Chem. A* **2009**, *113*, 5985.
- (17) James, T. D.; Sandanayake, K. R. A. S.; Shinkai, S. *Angew. Chem. Intl. Ed.* **1996**, *35*, 1910.
- (18) Zhu, L.; Shabbir, S. H.; Gray, M.; Lynch, V. M.; Sorey, S.; Anslyn, E. V. *J. Am. Chem. Soc.* **2006**, *128*, 1222.



- (19) Christinat, N.; Scopelliti, R.; Severin, K. *J. Org. Chem.* **2007**, *72*, 2192.
- (20) Welch, G. C.; Stephan, D. W. *J. Am. Chem. Soc.* **2007**, *129*, 1880.
- (21) Stephan, D. W. *Org. Biomol. Chem.* **2008**, *6*, 1535.
- (22) Spies, P.; Erker, G.; Kehr, G.; Bergander, K.; Fröhlich, R.; Grimme, S.; Stephan, D. W. *Chem. Commun.* **2007**, 5072.
- (23) Gille, A. L.; Gilbert, T. M. *J. Chem. Theory Comput.* **2008**, *4*, 1681.
- (24) Geier, S. J.; Gilbert, T. M.; Stephan, D. W. *J. Am. Chem. Soc.* **2008**, *130*, 12632.
- (25) Sumerin, V.; Schultz, F.; Nieger, M.; Atsumi, M.; Wang, C.; Leskela, M.; Pyykko, P.; Repo, T.; Rieger, B. *J. Organomet. Chem.* **2009**, *694*, 2654.
- (26) Mömning, C. M.; Otten, E.; Kehr, G.; Fröhlich, R.; Grimme, S.; Stephan, D. W.; Erker, G. *Angew. Chem., Int. Ed.* **2009**, *48*, 6643.
- (27) Rokob, T. A.; Hamza, A.; Pápai, I. *J. Am. Chem. Soc.* **2009**, *131*, 10701.
- (28) Hyla-Krispin, I.; Grimme, S. *Organometallics* **2004**, *23*, 5581.
- (29) Furche, F.; Perdew, J. P. *J. Chem. Phys.* **2006**, *124*, 044103.
- (30) Bühl, M.; Kabrede, H. *J. Chem. Theory Comput.* **2006**, *2*, 1282.
- (31) Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2007**, *127*, 124108.
- (32) Jiménez-Hoyos, C. A.; Janesko, B. G.; Scuseria, G. E. *J. Phys. Chem. A* **2009**, *113*, 11742.
- (33) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (34) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623–11627.
- (35) Staubitz, A.; Besora, M.; Harvey, J. N.; Manners, I. *Inorg. Chem.* **2008**, *47*, 5910.
- (36) Gilbert, T. M. *J. Phys. Chem. A* **2004**, *108*, 2550.
- (37) Lynch, B. J.; Fast, P. L.; Harris, M.; Truhlar, D. G. *J. Phys. Chem. A* **2000**, *104*, 4811–4815.
- (38) Phillips, J. A.; Cramer, C. J. *J. Chem. Theory Comput.* **2005**, *1*, 827.
- (39) Plumley, J. A.; Evanseck, J. D. *J. Phys. Chem. A* **2007**, *111*, 13472.
- (40) Plumley, J. A.; Evanseck, J. D. *J. Chem. Theory Comput.* **2008**, *4*, 1249.
- (41) Rakow, J. R.; Tüllmann, S.; Holthausen, M. C. *J. Phys. Chem. A* **2009**, *113*, 12035.
- (42) Svensson, M.; Humbel, S.; Froese, R. D. J.; Matsubara, T.; Sieber, S.; Morokuma, K. *J. Phys. Chem.* **1996**, *100*, 19357.
- (43) Janesko, B. G.; Henderson, T. M.; Scuseria, G. E. *Phys. Chem. Chem. Phys.* **2009**, *11*, 443.
- (44) Schwalbe, T.; Grimme, S. *Phys. Chem. Chem. Phys.* **2007**, *9*, 3397.
- (45) Johnson, E. R.; Becke, A. D.; Sherrill, C. D.; Di Labio, G. A. *J. Chem. Phys.* **2009**, *131*, 034111.
- (46) Wheeler, S. E.; Houk, K. N. *J. Chem. Theory Comput.* **2010**, *6*, 395.
- (47) Perdew, J. P.; Ruzsinszky, A.; Csonka, G. A.; Vydrov, O. A.; Scuseria, G. E.; Constantin, L. I.; Zhou, X.; Burke, K. *Phys. Rev. Lett.* **2008**, *100*, 136406.
- (48) Csonka, G. I.; Ruzsinszky, A.; Perdew, J. P.; Grimme, S. *J. Chem. Theory Comput.* **2008**, *4*, 888.
- (49) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868; **1997**, *78*, 1396(E).
- (50) Ahlrichs, R.; Penco, R.; Scoles, G. *Chem. Phys.* **1977**, *19*, 119.
- (51) Grimme, S.; Antony, J.; Schwabe, T.; Mück-Lichtenfeld, C. *Org. Biomol. Chem.* **2007**, *5*, 741.
- (52) Schwabe, T.; Grimme, S. *Acc. Chem. Res.* **2008**, *41*, 569.
- (53) Minenkov, Y.; Occhipinti, G.; Jensen, V. R. *J. Phys. Chem. A* **2009**, *113*, 11833.
- (54) *Gaussian Development Version*, Revision H.07; 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.; Parandekar, P. V.; Mayhall, N. J.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian, Inc.: Wallingford, CT, 2009.
- (55) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007.
- (56) Krishnan, R.; Binkley, J.; Seeger, R.; Pople, J. *J. Chem. Phys.* **1980**, *72*, 650.
- (57) Boys, S. F.; Bernardi, F. *Mol. Phys.* **1970**, *19*, 553.
- (58) Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200–1211.
- (59) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. *Phys. Rev. Lett.* **2003**, *91*, 146401.
- (60) Zhao, Y.; Truhlar, D. G. *J. Chem. Phys.* **2006**, *125*, 194101.
- (61) Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2006**, *110*, 13126.
- (62) Spies, P.; Fröhlich, R.; Kehr, G.; Erker, G.; Grimme, S. *Chem.—Eur. J.* **2008**, *14*, 333.
- (63) The Supporting Information of ref 22 considered the ground state of  $\text{Mes}_2\text{P}-\text{CH}_2-\text{CH}_2-\text{B}(\text{C}_6\text{F}_5)_2$ . This state has a B–P dative bond in a four-membered heterocycle. The reference reported energy differences relative to two higher-energy isomers with a broken B–P bond. Calculated energy differences between the ground state and the two bond-broken isomers were 6.8 and 7.0 kcal/mol with B97-D, and 11.3 and 10.4 kcal/mol with the more accurate spin-component-scaled<sup>75</sup> MP2 method.
- (64) Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2003**, *107*, 8996; **2004**, *108*, 1460(E).
- (65) Lynch, B. J.; Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2005**, *109*, 1643–1649.
- (66) Zheng, J.; Zhao, Y.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 808.
- (67) Iglesias, E.; Sordo, T. L.; Sordo, J. A. *Chem. Phys. Lett.* **1996**, *248*, 179.



- (68) Zhao, Y.; Lynch, B. J.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 2715–2719.
- (69) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Pople, J. A. *J. Chem. Phys.* **2000**, *112*, 7374–7382.
- (70) Zhao, Y.; Truhlar, D. A. *J. Chem. Theory Comput.* **2008**, *4*, 1849.
- (71) Antony, J.; Grimme, S. *Phys. Chem. Chem. Phys.* **2006**, *8*, 5287.
- (72) Vreven, T.; Morokuma, K. *J. Chem. Phys.* **1999**, *111*, 8799.
- (73) Curtiss, L. A.; Raghavachari, K. *Theor. Chem. Acc.* **2002**, *108*, 61.
- (74) The timings in Figure 3 are the total wall clock time in Link 502, divided by the number of SCF iterations. The calculations use “SCF=Tight, Integral(Grid=UltraFine)” and default Gaussian parameters otherwise.
- (75) Grimme, S. *J. Chem. Phys.* **2003**, *118*, 9095.

CT1000846

## Localized Orbitals for Incremental Evaluations of the Correlation Energy within the Domain-Specific Basis Set Approach

Joachim Friedrich\*

*Institute for Theoretical Chemistry, University of Cologne, Greinstr. 4,  
50939 Cologne, Germany*

Received February 19, 2010

**Abstract:** A modified version of the Boys localization method is proposed in order to make the domain-specific basis set approach in the framework of the incremental scheme (*J. Chem. Phys.* **2008**, *129*, 244105) generally applicable. The method optimizes the molecular orbitals in one atomic orbital basis set to be similar to localized molecular orbitals in a second atomic orbital basis set under the constraint that the molecular orbitals stay orthonormal. The procedure is tested for RI-MP2 incremental correlation energy expansions for aromatic systems like naphthalene, anthracene, and tetracene as well as for conjugated hydrocarbon chains like C<sub>20</sub>H<sub>2</sub>, C<sub>20</sub>H<sub>22</sub>, or *p*-quaterphenyle. For all investigated systems, a rapid convergence of the incrementally expanded correlation energies to the exact RI-MP2 energies is found. Furthermore, the systematic improvability of the approach is demonstrated.

### I. Introduction

Density functional theory (DFT) is today's most important quantum chemical method for applications to large systems. The price one has to pay for the increased range of applicability is a lack of systematical improvability. In contrast to DFT methods, wave-function-based correlation methods like many-body perturbation theory (MBPT), configuration interaction (CI), or coupled cluster (CC) are systematically improvable, but their unfavorable scaling with respect to the system size limits their application to small- or medium-sized molecules. The basic idea of local correlation methods is to overcome the unfavorable scaling behavior of the post Hartree–Fock methods by exploiting the local character of the electron correlation. During the past few decades, the development of local correlation methods was an active field in the quantum chemistry community.<sup>1–22</sup> Within the local domain approximation of Pulay and Saebø,<sup>3,23</sup> very efficient local versions of MP2,<sup>24,25</sup> CCSD,<sup>5</sup> and CCSD(T)<sup>26</sup> are available for molecular systems. The extension to periodic systems has been implemented in the CRYSCOR program<sup>27–29</sup> at the MP2 level of theory. Recently, Subotnik and co-workers proposed the use of

bump-functions to obtain smooth potential energy surfaces for this type of approach.<sup>6,30,31</sup> The extension of the Pulay approach to MRCI theory was recently proposed by Carter and co-workers.<sup>14,32</sup>

A conceptually different strategy to obtain a local correlation method is to divide the total system into small fragments and calculate the total energy on the basis of calculations of the small fragments. Within this category, the fragment molecular orbital approach,<sup>7,33</sup> the divide and conquer approach,<sup>10,11</sup> the cluster-in-molecule approach,<sup>22,34</sup> the systematic fragmentation method,<sup>12</sup> and the natural linear scaling coupled cluster<sup>9,35</sup> were proposed. Another fragment-based local correlation approach is the incremental scheme of Stoll.<sup>4,36,37</sup> It is a generalization of the Bethe–Goldstone expansion as introduced to the quantum chemistry community by Nesbet.<sup>1,38,39</sup> In an incremental calculation, the total correlation energy is expanded in a series of correlation energies of small domains.<sup>40–42</sup>

$$E_{\text{corr}} = \sum_{\substack{\mathbf{X} \\ \mathbf{X} \in \mathcal{P}(\mathbf{D}) \wedge |\mathbf{X}| \leq \varrho}} \Delta \varepsilon_{\mathbf{X}} \quad (1)$$

where  $\mathbf{X}$  is the summation index,  $\mathbf{D}$  is the set of domains,  $\mathcal{P}(\mathbf{D})$  is the power set of  $\mathbf{D}$ , and  $\varrho$  is the order of the expansion. The general increment  $\Delta \varepsilon_{\mathbf{X}}$  is defined as

\* E-mail: joachim\_friedrich@gmx.de.

$$\Delta\varepsilon_{\mathbb{X}} = \varepsilon_{\mathbb{X}} - \sum_{\substack{\mathbb{Y} \\ \mathbb{Y} \in \mathcal{P}(\mathbb{X}) \wedge |\mathbb{Y}| < |\mathbb{X}|}} \Delta\varepsilon_{\mathbb{Y}} \quad (2)$$

where  $\varepsilon_{\mathbb{X}}$  is the correlation energy of the domain  $\mathbb{X}$ . Since 1992, the incremental scheme was successfully applied to periodic systems<sup>43–48</sup> and finite systems.<sup>40–42,49–52</sup> Recently, the incremental scheme was also successfully applied to describe adsorption processes.<sup>53–56</sup> The drawback of the method so far was the large amount of hand work required to obtain a correlation energy according to eq 1. Due to the nature of the power set in eq 1, the number of calculations increases very rapidly if the number of domains increases. Since the higher-order increments become negligibly small if the distance of the one-site domains increases, one can safely neglect them without affecting the total accuracy of the calculation. In medium-sized molecules, the number of non-negligible increments is still on the order of 100, and an incremental calculation gets tedious for the one doing computations. To overcome this drawback, we proposed a fully automated implementation of the incremental scheme for molecular systems.<sup>51</sup> With this tool, we were able to investigate the performance of the incremental scheme for MP2, CCSD, CCSD(T), and RCCSD energies with respect to accuracy and efficiency.<sup>40–42,51,52,57,58</sup> Furthermore, the approach was extended to molecular dipole moments and quadrupole moments,<sup>59</sup> to treat the core–valence correlation in an efficient manner<sup>60</sup> and to explicitly correlated MP2 and CCSD theory.<sup>61</sup>

Recently, we proposed a domain-specific basis set approach for incremental evaluations of the coupled cluster singles and doubles and perturbative triples (CCSD(T)) energies.<sup>41,42</sup> It was demonstrated that the approach leads to a significant reduction of RAM and disk space requirements as well as CPU time. Due to the fact that the incremental scheme is inherently parallel, the computations of single increments were distributed over 20–50 nodes of a cheap cluster of standard PCs with a standard 100 megabit Ethernet connection. The key step in this approach is the reduction of the AO basis set in the incremental energy calculations. Since the domains are associated with a local region in space, one can divide the total AO basis set into two parts, the active part, which is spatially close to orbitals in the domain, and the environment, which is the rest of the system. Now we use the large original basis set in the active part and a small basis set in the environment. In order to obtain a set of local and orthogonal orbitals in the new basis set, we perform a Hartree–Fock (HF) calculation with a subsequent Boys localization.<sup>62</sup> The main problem associated with this procedure is the identification of the occupied orbitals of the domain in the new basis set. In the implementation in ref 42, the mapping of an occupied orbital  $\phi_a$  in the basis  $\mathcal{B}_1$  to the occupied orbital  $\phi'_a$  in the basis  $\mathcal{B}_2$  was accomplished by identifying their centers of charge  $\bar{R}_a$ :

$$\bar{R}_a(\mathcal{B}_1) \rightarrow \bar{R}_a(\mathcal{B}_2) \quad (3)$$

This procedure works very well if a unique maximum of the Boys functional exists, e.g., in  $\sigma$ -bonded hydrocarbons, water clusters, etc. For systems with more than one possible

localization maximum, this procedure fails, since the mapping in eq 3 is usually not fulfilled for all occupied orbitals of the system. In this work, a modified Boys localization procedure is implemented, where eq 3 holds for all localized orbitals of a system. The key step of the localization is to use the overlap of the molecular orbitals to a second local set of orbitals, which was previously introduced by Angeli et al.<sup>63</sup> as well as by Ahmadi and Røgggen.<sup>64</sup> Within the framework of the incremental scheme, the approach is tested for various critical systems at the RI-MP2 level using the TURBOMOLE program package.<sup>65</sup>

## II. Theory

The incremental scheme and the applied approximations were discussed in detail in refs 43 and 48–50. Therefore, we just give a brief introduction of the applied approximations.

**A. The Incremental Method.** Equation 1 combines the correlation energies of small subsystems in a systematic manner to obtain a controllable and systematically improvable approximation of the total correlation energy without calculating the correlation energy of the whole system. The starting point of such a calculation is localized HF orbitals obtained by a standard procedure like Boys<sup>62</sup> or Pipek–Mezey.<sup>66</sup> Since eq 1 requires disjoint subsets of occupied orbitals, the automatic domain generation of ref 51 is applied to accomplish this task. This method transforms the problem into a graph partitioning problem and applies a standard library routine<sup>67</sup> to obtain the desired domains.

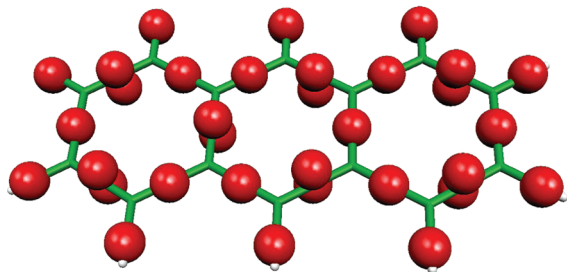
*1. Localized Orbitals and Perturbation Theory.* In order to use localized orbitals in combination with an unmodified MP2 code, one has to account for the fact that the Fock matrix is not diagonal in this basis. In the framework of the incremental scheme, this can be done by diagonalization of the Fock matrix in the active space of the domain. In this way, the incremental expansion ensures the canonical condition, since the higher orders correct for the nondiagonal Fock matrix.<sup>41,42,59</sup>

*2. Distance Screening.* The distance screening is an essential ingredient for an efficient incremental approach, since the number of calculations  $\mathcal{N}_{\text{calc}}$  grows very rapidly with increasing order  $\mathcal{O}$  and number of domains  $|\mathbb{D}|$ .

$$\mathcal{N}_{\text{calc}} = \sum_{i=1}^{\mathcal{O}} \binom{|\mathbb{D}|}{i} \quad (4)$$

Therefore, a straightforward application of eq 1 including all possible terms is usually not efficient with respect to CPU time. On the other hand, most of the increments are negligibly small, and one can neglect them without affecting the total accuracy of the calculation. The magnitude of an increment depends on the distances of the one-site domains as well as on the order of the increment (the number of one-site domains in the  $n$ -site domain).<sup>41,42</sup> Therefore, we introduced the order-dependent distance parameter to remove the negligible increments from the expansion:

$$t_{\text{dist}} = \frac{f}{(\mathcal{O} - 1)^2} \quad \mathcal{O} \geq 2$$



**Figure 1.** Centers of charge (spheres) for the Boys-localized occupied orbitals of anthracene.

where  $f$  is an adjustable parameter with typical values of about 30 Bohr.

3. *Domain-Specific Basis Set Approach.* The second ingredient for an efficient incremental approach is the domain-specific basis. The basic idea of this approximation is the fact that virtual orbitals far from the domain do not significantly contribute to the correlation energy of an arbitrary domain. Therefore, one can use a smaller basis set for the environment of a domain.<sup>41,42,68</sup> For a systematic choice of the basis set, we use a sphere with the radius  $t_{\text{main}}$  for every occupied orbital associated with a given  $n$ -site domain (labeled with  $\mathbb{X}$ ). This means we map a set of atoms to every occupied orbital of the domain:

$$\phi_i \rightarrow \{\bar{r}_{i_1}, \bar{r}_{i_2}, \dots\} = A_{\phi_i} \quad (5)$$

where the  $\bar{r}$  are coordinates of atoms in the molecule. The active part  $A_{\mathbb{X}}$  of an  $n$ -site domain is obtained by unifying the sets associated with the orbitals of the domain:

$$A_{\mathbb{X}} = \bigcup_{\phi_i \in \mathbb{X}'} A_{\phi_i} \quad \mathbb{X}' = \bigcup_{X \in \mathbb{X}} X \quad (6)$$

where the sets  $\mathbb{X}'$  have to be introduced formally to account for the fact that  $n$ -site domains are sets of sets of occupied orbitals. Now we use the large original basis for all atoms in  $A_{\mathbb{X}}$ , and the rest of the molecule is treated with the smaller basis set. In order to obtain local orthogonal orbitals, a HF calculation with a subsequent localization is performed (vide infra). Besides the reduction of CPU time, the domain-specific basis set approach reduces the number of two-electron integrals significantly and therefore the disk and RAM space requirements.

The key step in this approach is the mapping of a set of local occupied orbitals in one basis to a set of local occupied orbitals in another basis set (eq 3). This mapping step is problematic, since in many real life molecules the set of local orbitals is not unique for symmetry reasons. This can be easily demonstrated using the centers of charge of anthracene (Figure 1): In the ring systems, one can identify the alternating single and double bonds as usually drawn in the Lewis structure. Considering another resonance structure, one can immediately see that there is another equivalent choice for the centers of charge, where the Boys functional has a maximum. The Boys functional just maximizes the distances of the centers of charges, and both localization maxima are equivalent due to symmetry. Therefore, it is usually not predictable to which extremum the Boys localization converges.

**B. Template Localization.** The standard Boys localization procedure does not necessarily yield occupied orbitals which are sufficiently similar (vide supra), if different AO basis sets are used in a calculation. Therefore, we impose a further condition to accomplish this requirement. The starting point of the procedure is a set of local occupied orbitals (template orbitals) in a small AO basis set  $\mathcal{B}_2$ . The idea is now to make the occupied HF orbitals in a second AO basis  $\mathcal{B}_1$  as equal as possible to the template orbitals without affecting their orthogonality. This can be accomplished by maximizing the functional:

$$D = \sum_k \langle \phi_k^{\mathcal{B}_1} | \phi_k^{\mathcal{B}_2} \rangle \quad (7)$$

where  $\phi_k^{\mathcal{B}_1}$  refers to molecular orbitals in the basis  $\mathcal{B}_1$  and  $\phi_k^{\mathcal{B}_2}$  represents orbitals in the basis  $\mathcal{B}_2$ . Note that conceptually similar functionals were proposed by Angeli et al.<sup>63</sup> and Ahmadi and Røgggen.<sup>64</sup> A maximization of the functional in eq 7 leads to a set of target functions in the basis  $\mathcal{B}_1$  which are similar to the template functions in the basis  $\mathcal{B}_2$ . In order to preserve the orthogonality, we use orthogonal  $2 \times 2$  rotations to perform the transformations. This means the matrices  $O^{ij}$  which mix the orbitals  $\phi_i$  and  $\phi_j$  have the form

$$O^{ij} = \begin{pmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{pmatrix}^{ij} \quad (8)$$

or

$$O^{ij} = \begin{pmatrix} -\cos(\alpha) & \sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{pmatrix}^{ij} \quad (9)$$

where  $\alpha$  is the rotation angle. In contrast to the Edmiston–Ruedenberg optimization scheme,<sup>69</sup> where only  $2 \times 2$  rotations of the type in eq 8 are applied, we need more flexibility with the second type of rotations in eq 9. A simple example for the need of the second type of rotations is the interchange of two orbitals: We assume that the second set of orbitals is equal to the first set of orbitals, except for a swap of two orbitals. The optimal step is now to swap the two orbitals back, without changing the sign, which corresponds to the second type of rotations. This can be seen explicitly in eq 16 and eq 19 if the corresponding matrix elements are calculated. Equation 19 will find the rotation, whereas eq 16 does not. Since the optimization maximizes the value of  $D$ , we have to deal with the second type of rotations in order to include all possible orbital transformations.

With the first matrix, the orbitals read

$$\begin{aligned} \langle \tilde{\phi}_i^{\mathcal{B}_1} | &= \cos(\alpha) \langle \phi_i^{\mathcal{B}_1} | + \sin(\alpha) \langle \phi_j^{\mathcal{B}_1} | \\ \langle \tilde{\phi}_j^{\mathcal{B}_1} | &= -\sin(\alpha) \langle \phi_i^{\mathcal{B}_1} | + \cos(\alpha) \langle \phi_j^{\mathcal{B}_1} | \end{aligned} \quad (10)$$

Inserting the ansatz into eq 7, we obtain

$$\begin{aligned} D_{ij}(\alpha) &= \sum_{k \neq i,j} \langle \phi_k^{\mathcal{B}_1} | \phi_k^{\mathcal{B}_2} \rangle + \cos(\alpha) [\langle \phi_i^{\mathcal{B}_1} | \phi_i^{\mathcal{B}_2} \rangle + \langle \phi_j^{\mathcal{B}_1} | \phi_j^{\mathcal{B}_2} \rangle] \\ &\quad + \sin(\alpha) [\langle \phi_j^{\mathcal{B}_1} | \phi_i^{\mathcal{B}_2} \rangle - \langle \phi_i^{\mathcal{B}_1} | \phi_j^{\mathcal{B}_2} \rangle] \\ &= \sum_{k \neq i,j} \langle \phi_k^{\mathcal{B}_1} | \phi_k^{\mathcal{B}_2} \rangle + \cos(\alpha) A_{ij} + \sin(\alpha) B_{ij} \end{aligned} \quad (11)$$

Now, we define  $\beta$  by



$$\tan \beta = \frac{B_{ij}}{A_{ij}} \quad (12)$$

Using this definition, we arrive after some algebraic manipulations at

$$D_{ij}(\alpha) = \sum_{k \neq i,j} \langle \phi_k^{\beta_1} | \phi_k^{\beta_2} \rangle + \sqrt{A_{ij}^2 + B_{ij}^2} \cos(\alpha - \beta) \quad (13)$$

Since the prefactor of the cosine is always positive, the functional in eq 13 is maximal if the cosine is +1 and minimal if the cosine is -1. This is fulfilled if  $\alpha - \beta = 0$  and  $\alpha - \beta = \pi$ , and thus we get

$$\alpha^{\max} = \beta \quad \alpha^{\min} = \beta + \pi \quad (14)$$

Explicitly, the angle  $\alpha_{\max}$  is calculated as

$$\alpha_{ij}^{\max} = \arccos\left(\frac{A_{ij}}{\sqrt{A_{ij}^2 + B_{ij}^2}}\right) \quad (15)$$

Up to now, we found the maximal increase for  $D(\alpha)$  for a given pair of functions  $\langle \phi_i^{\beta_1} | \phi_j^{\beta_1} \rangle$ . In order to find the maximum increase of  $D(\alpha)$  with respect to the choice of all possible orbital pairs  $i, j$ , we use the matrix  $\mathbf{D}^{\max}$  with the entries

$$\mathbf{D}_{ij}^{\max} = [\mathbf{D}_{ij}^{\max}(\alpha) - D] = -A_{ij} + \sqrt{A_{ij}^2 + B_{ij}^2} \quad (16)$$

The difference between the  $\mathbf{D}_{ij}^{\max}(\alpha)$  and  $D$  yields the increase of the functional  $D$  with respect to a  $2 \times 2$  rotation of the orbitals  $i, j$ . Therefore, the matrix  $\mathbf{D}^{\max}$  contains all possible changes. Note that we do not have a dependence on the rotation angle  $\alpha$ , since we insert the maximal increase for every orbital pair.

Before we proceed with the final optimization step, we have to consider the second type of rotation in eq 9. In this case, the orbitals read

$$\begin{aligned} \langle \tilde{\phi}_i^{\beta_1} | &= -\cos(\alpha') \langle \phi_i^{\beta_1} | + \sin(\alpha') \langle \phi_j^{\beta_1} | \\ \langle \tilde{\phi}_j^{\beta_1} | &= \sin(\alpha') \langle \phi_i^{\beta_1} | + \cos(\alpha') \langle \phi_j^{\beta_1} | \end{aligned} \quad (17)$$

In the further calculation, we obtain basically the same equations as above. The only difference is the definition of  $A_{ij}$  and  $B_{ij}$  in the square root:

$$\begin{aligned} A'_{ij} &= -\langle \phi_i^{\beta_1} | \phi_i^{\beta_2} \rangle + \langle \phi_j^{\beta_1} | \phi_j^{\beta_2} \rangle \\ B'_{ij} &= \langle \phi_j^{\beta_1} | \phi_i^{\beta_2} \rangle + \langle \phi_i^{\beta_1} | \phi_j^{\beta_2} \rangle \end{aligned} \quad (18)$$

where the prime is used to indicate rotations of the second type. For the second type of rotations, eq 16 reads

$$\mathbf{D}_{ij}^{\max} = [\mathbf{D}_{ij}^{\max}(\alpha) - D] = -A'_{ij} + \sqrt{A'_{ij}{}^2 + B'_{ij}{}^2} \quad (19)$$

The final optimization setup is as follows: We search the largest value in  $\mathbf{D}^{\max}$  and  $\mathbf{D}'^{\max}$  to find the orbital pair with the largest increase of  $D$  and perform the associated  $2 \times 2$  rotation. This is repeated, until all off diagonal elements in  $\mathbf{D}^{\max}$  and  $\mathbf{D}'^{\max}$  are lower than a given threshold.

The straightforward application of the procedure above might lead to delocalized orbitals, if the optimization

**Table 1.** List of the Applied Truncation Parameters

dsp	domain size parameter; a rough measure for the size of the domains <sup>51</sup>
$t_{\text{con}}$	connectivity parameter; sets the connectivity for far distant orbitals to zero <sup>51</sup>
$t_{\text{main}}$	the radius around active orbitals to determine the basis for the individual calculations (section II.A.3, ref 41)
core	number of frozen core orbitals
$f$	parameter for the order dependent distance screening using $\#(\varrho - 1)^2$ (section II.A.2, refs 41, 42)

procedure ends in a local maximum. The reason for this is that the locality of the orbitals comes only implicitly due to the locality of the template orbitals. To overcome this problem, we use Boys orbitals as an initial guess, apply the procedure above, and finally perform a second Boys localization at the end. The first two orthogonal transformations create a set of local orbitals with charge centers close to the template functions. The final Boys localization ensures that the orbitals fulfill an explicit localization criterion; i.e., the Boys functional is maximal.

The composition of these three orthogonal transformations leads to a sufficiently stable algorithm to perform incremental calculations within the domain-specific basis set approach. Note that the orbitals of the composite transformation are not equivalent to Boys orbitals obtained by a standard one-step localization.

### III. Computational Details

If nothing else is stated, the geometries were optimized at the RI-BP86/TZVP level of theory using the TURBOMOLE quantum chemistry program package.<sup>65,70-74</sup> Stationary points were characterized by analyzing the Hessian matrix.<sup>75</sup> The RI-MP2 energies were computed with the ricc2 module<sup>76</sup> of TURBOMOLE 5.10. The necessary data such as the Fock matrix, the localization matrix, dipole integrals, as well as the overlap matrix of two different AO basis sets were obtained by an interface to a development version of TURBOMOLE.

**A. Incremental Calculations.** The threshold for the maximum values in the matrices  $\mathbf{D}^{\max}$  and  $\mathbf{D}'^{\max}$  was set to  $10^{-8}$  in the template localization step, whereas the Boys localizations were converged to  $10^{-11}$ . The occupied orbitals in the two basis sets were identified by their centers of charge with an identification tolerance of 0.1 Bohr. In the incremental energy evaluations, the RI basis (cbas) of the large original basis set was used for the total system. To get an overview of the applied truncation parameters in our incremental calculations, we included a list of the parameters with a short description (Table 1). In the environment, the SVP basis was applied for all systems in this study.

### IV. Applications

A critical test for the stability of the proposed procedure is the evaluation of increments within the domain-specific basis set approach. First, the incremental scheme requires a large number of calculations; e.g., in order to get a single correlation energy, it is not unusual that a few hundred correlation calculations are necessary. In the domain-specific

**Table 2.** Convergence of the Incremental RI-MP2 Correlation Energies<sup>a</sup>

order	C <sub>10</sub> H <sub>22</sub>			C <sub>20</sub> H <sub>42</sub>			C <sub>20</sub> H <sub>22</sub>		
	$E_{\text{corr}}(l)$ [a.u.]	error [kcal/mol]	$E_{\text{corr}}$ [%]	$E_{\text{corr}}(l)$ [a.u.]	error [kcal/mol]	$E_{\text{corr}}$ [%]	$E_{\text{corr}}(l)$ [a.u.]	error [kcal/mol]	$E_{\text{corr}}$ [%]
TZVP									
1	-1.188660	209.80	78.05	-2.293514	461.38	75.72	-1.995851	489.70	71.89
2	-1.520772	1.39	99.85	-3.021870	4.33	99.77	-2.761895	9.00	99.48
3	-1.523373	-0.24	100.03	-3.030365	-1.00	100.05	-2.775034	0.75	99.96
exact	-1.522990			-3.028767			-2.776232		
QZVP									
1	-1.464292	254.13	78.34	-2.826150	558.80	76.04	-2.460658	587.39	72.44
2	-1.867051	1.39	99.88	-3.710525	3.84	99.84	-3.381958	9.26	99.57
3	-1.869324	-0.04	100.00	-3.716863	-0.14	100.01	-3.395008	1.08	99.95
exact	-1.869267			-3.716647			-3.396722		

<sup>a</sup> core = 10 (C<sub>10</sub>H<sub>22</sub>), core = 20 (C<sub>20</sub>H<sub>42</sub>, C<sub>20</sub>H<sub>22</sub>); dsp = 4;  $t_{\text{main}}$  = 3 Bohr;  $f$  = 25 Bohr (C<sub>10</sub>H<sub>22</sub>, C<sub>20</sub>H<sub>42</sub>),  $f$  = 30 Bohr (C<sub>20</sub>H<sub>22</sub>);  $t_{\text{con}}$  = 3 Bohr.

basis set, this means that the localization has to work for every single calculation. Second, the structure of the basis for the calculation in the  $n$ -site domains is more complicated, since there might be several regions with different basis sets in the molecule; e.g., in higher order domains, the orbital domains are not necessarily local anymore. In cases where the Boys functional has several symmetry equivalent maxima like in benzene, it is not predictable and not controllable to which maximum the Boys localization will converge. With the proposed algorithm, we were able to overcome this drawback for all cases tested so far.

The main goal of this work is the test of the potential accuracy of the domain-specific basis set approach in combination with the template localization. Within the efficient RI-MP2 routines in TURBOMOLE,<sup>76,77</sup> there is no need to make local approximations in the correlation part for the molecules in this study, since the HF calculation consumes a large part of the CPU time in our calculations. Clearly, this will change if coupled cluster methods are used. However, to test the performance of the approach with respect to the accuracy on a large set of molecules, we decided to use MP2, since when the domain-specific basis set approach is used the convergence of the MP2 energies is similar to the convergence of the coupled cluster energies.<sup>41,42,68</sup>

**A. Boys Systems.** The first issue to study is the performance of the new approach for systems where no problems with the ambiguity of the localization exist, i.e., Boys systems, in order to investigate how the proposed procedure might affect the accuracy in these cases.

*1. Hydrocarbons.* An easy test case for local correlation methods is the use of unbranched hydrocarbon chains. In Table 2, we present the results for decane, eicosane, and the unsaturated C<sub>20</sub>H<sub>22</sub> in the TZVP and in the QZVP basis sets, respectively. For both saturated hydrocarbons in the TZVP basis set, the convergence of the incremental series is fast, and a third-order expansion is sufficient to obtain chemical accuracy of about 1 kcal/mol. The relative correlation energy is 100.03% and 100.05% for decane and eicosane, respectively. Increasing the basis set to the quadruple- $\zeta$  level slightly improves the accuracy of the incrementally expanded energy. At third-order level, the errors are -0.04 and -0.14 kcal/mol.

**Table 3.** Convergence of the Incremental RI-MP2 Correlation Energies of the (H<sub>2</sub>O)<sub>13</sub> Cluster in Figure 2 Using the aug-cc-pVXZ Basis Set Series of Dunning and Co-Workers<sup>79,80 a</sup>

order	$E_{\text{corr}}(l)$ [a.u.]	error [kcal/mol]	$E_{\text{corr}}$ [%]
aug-cc-pVDZ			
1	-2.824690	72.96	96.05
2	-2.942911	-1.22	100.07
3	-2.940631	0.21	99.99
exact	-2.940959		
aug-cc-pVTZ			
1	-3.456064	79.23	96.48
2	-3.583812	-0.94	100.04
3	-3.582025	0.19	99.99
exact	-3.582320		
aug-cc-pVQZ			
1	-3.682562	78.49	96.72
2	-3.808779	-0.71	100.03
3	-3.807413	0.14	99.99
exact	-3.807641		

<sup>a</sup> core = 13, dsp = 4 Bohr,  $t_{\text{main}}$  = 3 Bohr,  $f$  = 25 Bohr,  $t_{\text{con}}$  = 3 Bohr.

More difficult examples for local correlation methods are conjugated  $\pi$  systems. Therefore, we study C<sub>20</sub>H<sub>22</sub> with alternating single and double bonds. Using the same truncation parameters as for the saturated hydrocarbons, the errors for the unsaturated C<sub>20</sub> chain are 0.75 kcal/mol in the TZVP basis set and 1.08 kcal/mol in the QZVP basis set. Comparing the accuracy of the RI-MP2 correlation energies for the saturated hydrocarbons and the unsaturated hydrocarbon, we find a slightly higher accuracy for saturated hydrocarbons. We note that the accuracy can be increased by increasing the domain size parameter (dsp) or the radius for the basis set truncation ( $t_{\text{main}}$ ).

*2. Water Cluster.* The incremental scheme yields very accurate results for molecular clusters. With the domain-specific basis set approach, it was demonstrated earlier that the incremental scheme can be applied very efficiently for water clusters in the framework of coupled cluster.<sup>42,68</sup> Table 3 shows the convergence of the incremental RI-MP2 correlation energies with respect to the exact RI-MP2 correlation energy for the aug-cc-pVXZ (X = D,T,Q) basis set series of Dunning et al.<sup>79,80</sup> The accuracy of the incremental expansions is similar for the applied aug-cc-pVXZ basis sets. At second order, the errors are on the order of 1 kcal/mol,

**Table 4.** Dependence of RI-MP2 Correlation Energies, Errors, and Relative Correlation Energies of  $C_{20}H_2$  on the Truncation Parameters of Table 1<sup>a</sup>

entry no.	order	dsp	$f$ [Bohr]	$t_{\text{main}}$ [Bohr]	$E_{\text{corr}}(l)$ [a.u.]	error [kcal/mol]	$E_{\text{corr}}$ [%]
basis = TZVP							
1	3	4	25	3	-2.640592	3.15	99.81
2	3	4	30	3	-2.643350	1.42	99.91
3	4	4	30	3	-2.644079	0.96	99.94
4	3	4	40	3	-2.643461	1.35	99.92
5	4	4	40	3	-2.644190	0.89	99.95
6	3	5	25	3	-2.643765	1.15	99.93
7	3	6	25	3	-2.644735	0.55	99.97
8	3	4	25	5	-2.640372	3.28	99.80
9	4	4	25	5	-2.643584	1.27	99.92
10	3	4	30	5	-2.643920	1.06	99.94
11	4	4	30	5	-2.644121	0.93	99.94
12	3	4	40	5	-2.643965	1.03	99.94
13	4	4	40	5	-2.644167	0.90	99.95
14	3	5	40	5	-2.644633	0.61	99.96
15	4	5	40	5	-2.645062	0.34	99.98
exact					2.645605		
basis = QZVP							
16	3	5	40	5	-3.201505	0.73	99.96
exact					-3.202670		

<sup>a</sup> core = 20,  $t_{\text{con}} = 3$  Bohr.

and at third order, they are below 1 kcal/mol (0.21, 0.19, 0.14 kcal/mol).

Finally, we conclude that the convergence of the incremental series in the domain-specific basis-set approach using template localized orbitals is sufficiently fast, and chemically accurate results can be obtained for the Boys systems.

**B. Non-Boys-Systems.** The focus of this work is to demonstrate the performance of the template localization for the cases where the simple procedure outlined in ref 42 does not work, i.e., non-Boys systems. Therefore, we applied the approach to various systems of chemical interest such as conjugated  $\pi$  systems and aromatic compounds with a highly delocalized nature.

1.  $C_{20}H_2$ . As a first critical test system, we applied the domain-specific basis set approach in combination with the template localization to  $C_{20}H_2$ . This molecule is a challenge for local correlation methods since it has alternating single and triple bonds. Furthermore, the treatment of triple bonds is not possible within the domain-specific basis set approach, if standard Boys orbitals are used. The problems occur in the identification of the localized orbitals within two different basis sets, since the localization is not unique for triple bonds. For example, maximizing the distances of the charge centers in a triple bond results in a triangle, which can be rotated around the bond axis without changing the Boys functional. Practically, this means that the program terminates with an error since

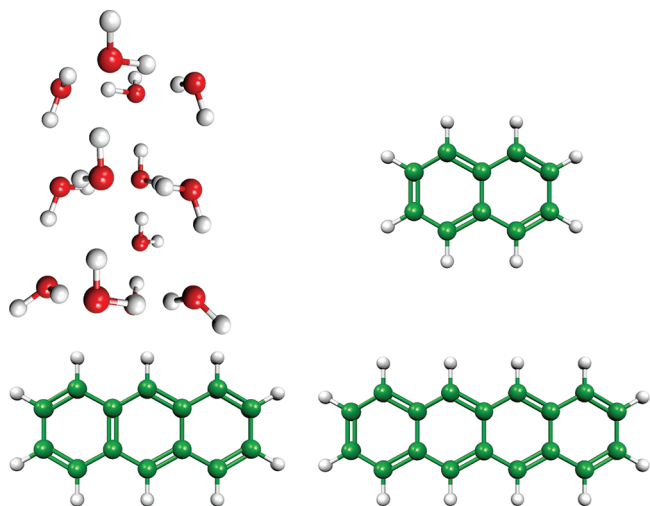
$$|\vec{R}_a(\mathcal{B}_1) - \vec{R}_a(\mathcal{B}_2)| < 0.1 \text{ Bohr}$$

is usually not fulfilled for Boys orbitals. With the proposed localization scheme, the identification of the occupied orbitals within two different basis sets could be done for all molecules and for all increments in this study. The performance of the incremental scheme in combination with template-localized orbitals is presented in Table 4 for various truncation

parameters (order, dsp,  $f$ , and  $t_{\text{main}}$ ). Using dsp = 4,  $f = 25$ , and  $t_{\text{main}} = 3$  as for the saturated hydrocarbon compounds above, the error of the RI-MP2/TZVP correlation energy is 3.15 kcal/mol at third order. Increasing the order-dependent distance threshold  $f$  from 25 to 30 Bohr, the error decreases by about a factor of 2 to 1.42 kcal/mol. A further increase of  $f$  to 40 Bohr does not significantly improve the energy (entry 4). The error at fourth order is 0.96 kcal/mol for  $f = 30$  and 0.89 kcal/mol for  $f = 40$  (entries 3 and 5). This small change in the energies with respect to the distance threshold  $f$  indicates that the significant increments are included already for  $f = 30$ . This observation is equivalently true when comparing the entry pairs 10 and 12 and 11 and 13 with a larger  $t_{\text{main}}$ . Increasing the domain size from dsp = 4 to dsp = 6 decreases the error to 0.55 kcal/mol (entry 7). Increasing the radius for the basis set truncation from  $t_{\text{main}} = 3$  to  $t_{\text{main}} = 5$  decreases the error to 1.06 kcal/mol for the third order calculation (entry 10). At the fourth order level, the error is 0.93 kcal/mol (entry 11). For dsp = 5 and  $t_{\text{main}} = 5$ , the error is 0.61 kcal/mol at third order and 0.34 kcal/mol at fourth order (entries 14 and 15). Comparing entries 1 and 8, we find a slightly smaller fraction of the correlation energy for the calculation with the larger  $t_{\text{main}}$  in entry 8. This behavior can be explained by the tight distance truncation with  $f = 25$  which causes an error around 2 kcal/mol. If we compare the corresponding entry pairs with  $f = 30$  (entries 2 and 10), we find that the increase of  $t_{\text{main}}$  from 3 to 5 yields again a higher accuracy. These findings can be explained by the fact that a larger  $t_{\text{main}}$  can lead to larger contributions of the individual increments. If the distance truncation is as serious as in entries 1 and 8, it is not surprising that a larger  $t_{\text{main}}$  does not improve the total accuracy, since the sum of the neglected increments is large for  $f = 25$  and it was slightly increased by the increase of  $t_{\text{main}}$ . Increasing the basis set from TZVP to QZVP increases the error slightly from 0.61 to 0.73 kcal/mol (entry 16). Note that the change in the total energy is ca. 350 kcal/mol in going from TZVP to QZVP.

From these findings, it is evident that the truncation parameters can be used to control the accuracy of the incremental scheme in a systematic manner and that the template localization works sufficiently well for this difficult system.

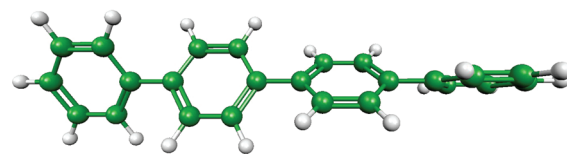
2. *Polycyclic Aromatic Hydrocarbons.* Next, the proposed localization procedure is tested for polycyclic aromatic hydrocarbons (Figure 2). Therefore, we check the performance of the approach for naphthalene, anthracene, and tetracene (Table 5). Considering the accuracy of the incremental series, we obtain errors of about 1 kcal/mol for a third-order calculation using the same thresholds as for the saturated hydrocarbons above. The relative correlation energy ranges from 99.97% to 100.05% in the TZVP basis set. If the larger QZVP basis set is used, the errors increase slightly to -0.88 and -1.09 kcal/mol for naphthalene and anthracene, respectively. Due to the accuracy of the results, we conclude that polycyclic aromatic hydrocarbons can be treated with the incremental scheme. As expected for such highly delocalized systems, they are slightly more difficult to treat than their saturated counterparts.



**Figure 2.** RI-BP86/TZVP optimized structures of naphthalene, anthracene, and tetracene. The geometry of the  $(\text{H}_2\text{O})_{13}$  cluster was taken from ref 78.

A further interesting test system for the localization method is the *p*-quaterphenyle molecule in Figure 3. The results of the third order calculations for different truncation parameters are given in Table 6. In this case, the results clearly indicate that one should not use the same truncation parameters as for the saturated hydrocarbons. On the other hand, the quality of the third order energies can be improved by increasing the domain sizes and the radius for the basis set truncation. With sufficiently large values for  $d_{\text{sp}}$  and  $t_{\text{main}}$ , one can obtain 99.94% of the correlation energy. The error is 1.33 kcal/mol, which is slightly above the desired error of 1 kcal/mol. However, this is a critical molecule for local correlation approaches, and the accuracy is still reasonable.

3. *Oligopeptide.* As a final example, we included the oligopeptide in Figure 4. For peptides without aromatic groups, the domain-specific basis set approach can be applied in combination with Boys orbitals as demonstrated in ref 41. To obtain a critical test of the proposed localization procedure, we included histidine, tryptophan, and phenylalanine. The convergence of the incremental series for the RI-MP2/TZVP correlation energy for this system is given in Table 7 using two sets of truncation parameters. The convergence of the incremental series is fast for both parameter sets since 99.89% and 99.95% of the correlation

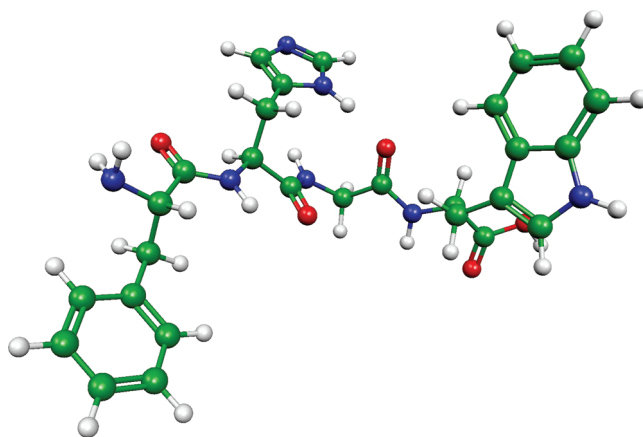


**Figure 3.** RI-BP86/TZVP optimized structure of *p*-quaterphenyle.

**Table 6.** Dependence of RI-MP2/TZVP Correlation Energies, Errors and Relative Correlation Energies of *p*-Quaterphenyle on the Parameters of Table 1<sup>a</sup>

order	dsp	$f$ [Bohr]	$t_{\text{main}}$ [Bohr]	$E_{\text{corr}}(l)$ [a.u.]	error [kcal/mol]	$E_{\text{corr}}$ [%]
3	4	$\infty$	3	-3.303295	-17.19	100.84
3	5	$\infty$	6	-3.256601	12.11	99.41
3	6	35	6	-3.273790	1.33	99.94
exact				-3.275906		

<sup>a</sup> core = 24,  $t_{\text{con}} = 3$  Bohr.



**Figure 4.** RI-BP86/SV(P) optimized structure of an oligopeptide.

energy are recovered at third order. At fourth order, 99.95% and 99.99% of the correlation energy are recovered. The absolute errors are still above 1 kcal/mol except for the fourth order calculation in combination with  $d_{\text{sp}} = 6$  and  $t_{\text{main}} = 5$  Bohr. The reason for the larger absolute error is the magnitude of the correlation energy.

Since the convergence of the MP2 correlation energies in the domain-specific basis set approach is similar to the convergence of the corresponding coupled cluster energies

**Table 5.** Convergence of the Incremental RI-MP2 Correlation Energies for Naphthalene, Anthracene, and Tetracene<sup>a</sup>

order	naphthalene			anthracene			tetracene		
	$E_{\text{corr}}(l)$ [a.u.]	error [kcal/mol]	$E_{\text{corr}}$ [%]	$E_{\text{corr}}(l)$ [a.u.]	error [kcal/mol]	$E_{\text{corr}}$ [%]	$E_{\text{corr}}(l)$ [a.u.]	error [kcal/mol]	$E_{\text{corr}}$ [%]
TZVP									
1	-0.904435	284.87	66.58	-1.217251	426.96	64.14	-1.401001	651.19	57.45
2	-1.350451	5.00	99.41	-1.879245	11.55	99.03	-2.407246	19.76	98.71
3	-1.359066	-0.41	100.05	-1.898451	-0.50	100.04	-2.437892	0.53	99.97
exact	-1.358411			-1.897658			-2.438732		
QZVP									
1	-1.118346	342.97	67.17	-1.506995	512.63	64.85			
2	-1.655905	5.65	99.46	-2.303540	12.79	99.12			
3	-1.666305	-0.88	100.08	-2.325659	-1.09	100.07			
exact	-1.664910			-2.323917					

<sup>a</sup> Naphthalene: core = 10,  $d_{\text{sp}} = 4$ ,  $t_{\text{main}} = 3$  Bohr,  $f = \text{inf}$  Bohr,  $t_{\text{con}} = 3$  Bohr. Anthracene: core = 14,  $d_{\text{sp}} = 4$ ,  $t_{\text{main}} = 3$  Bohr,  $f = 30$  Bohr,  $t_{\text{con}} = 3$  Bohr. Tetracene: core = 18,  $d_{\text{sp}} = 4$ ,  $t_{\text{main}} = 3$  Bohr,  $f = 30$  Bohr,  $t_{\text{con}} = 3$  Bohr.



**Table 7.** Convergence of the Incremental RI-MP2/TZVP Correlation Energies for the Oligopeptide in Figure 4<sup>a</sup>

order	$E_{\text{corr}}(l)$ [a.u.]	error [kcal/mol]	$E_{\text{corr}}$ [%]
dsp = 4, $t_{\text{main}} = 3$ Bohr			
1	-4.609373	944.28	75.39
2	-6.103895	6.45	99.83
3	-6.107536	4.17	99.89
4	-6.111360	1.77	99.95
dsp = 6, $t_{\text{main}} = 5$ Bohr			
1	-4.758638	850.61	77.83
2	-6.109521	2.92	99.92
3	-6.111255	1.83	99.95
4	-6.113515	0.41	99.99
exact	-6.114175		

<sup>a</sup> core = 40, dsp=4,  $t_{\text{main}} = 3$  Bohr,  $f = 30$  Bohr,  $t_{\text{con}} = 3$  Bohr.

it seems to be a promising goal to use template localized orbitals at the coupled cluster level. We note that a significant reduction of the total CPU time was achieved with the domain-specific basis set approach at the CCSD(T) level.<sup>41</sup>

## V. Conclusion

We implemented a modified version of the Boys localization which can be applied to extend the domain-specific basis set approach to aromatic systems, where the original approach does not work. It was shown for aromatic systems like naphthalene, anthracene, and tetracene and conjugated hydrocarbon chains like C<sub>20</sub>H<sub>2</sub>, C<sub>20</sub>H<sub>22</sub>, and *p*-quaterphenyle that the localization procedure works sufficiently well. The accuracy of the incremental RI-MP2 correlation energies is close to a chemical accuracy of 1 kcal/mol for these difficult systems, if appropriate truncation parameters are used. Furthermore, it has been demonstrated that increasing the domain sizes or increasing  $t_{\text{main}}$  systematically improves the accuracy of the calculation. In the future, we plan to combine the template localization with the CCSD(T) implementation of the incremental scheme, to obtain a generally applicable systematically improvable and efficient incremental CCSD(T) method.

**Acknowledgment.** This work was supported by the German Research Foundation (DFG) through priority program 1145 and SFB 624. The author would like to acknowledge Prof. M. Dolg and Dr. A. Engels-Putzka for various discussions and carefully reading the manuscript, Prof. T. Helgaker for discussing the localization procedure, Dr. D. P. Tew for the required data interfaces to TURBO-MOLE, and T. Kjergaard for an interface to the required overlap integrals in DALTON.

## References

- Nesbet, R. K. *Phys. Rev.* **1967**, *155*, 51.
- Förner, W.; Ladik, J.; Otto, P.; Cizek, J. *Chem. Phys.* **1985**, *97*, 251.
- Pulay, P.; Saebø, S. *Theor. Chim. Acta.* **1986**, *69*, 357.
- Stoll, H. *Chem. Phys. Lett.* **1992**, *191*, 548.
- Hampel, C.; Werner, H.-J. *J. Chem. Phys.* **1996**, *104*, 6286.
- Subotnik, J. E.; Head-Gordon, M. *J. Chem. Phys.* **2005**, *123*, 64108.
- Fedorov, D. G.; Kitaura, K. *J. Chem. Phys.* **2004**, *121*, 2483.
- 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.
- Flocke, N.; Bartlett, R. J. *J. Chem. Phys.* **2004**, *121*, 10935.
- Kobayashi, M.; Imamura, Y.; Nakai, H. *J. Chem. Phys.* **2007**, *127*, 074103.
- Kobayashi, M.; Nakai, H. *J. Chem. Phys.* **2008**, *129*, 044103.
- Deev, V.; Collins, M. A. *J. Chem. Phys.* **2005**, *122*, 154102.
- Li, S.; Shen, J.; Li, W.; Jiang, Y. *J. Chem. Phys.* **2006**, *125*, 074109.
- Walter, D.; Szilva, A. B.; Niedfeld, K.; Carter, E. A. *J. Chem. Phys.* **2002**, *117*, 1982.
- Auer, A. A.; Nooijen, M. *J. Chem. Phys.* **2006**, *125*, 024104.
- Weijo, V.; Manninen, P.; Jørgenson, P.; Christiansen, O.; Olsen, J. *J. Chem. Phys.* **2007**, *127*, 074106.
- Doser, B.; Lambrecht, D. S.; Ochsenfeld, C. *Phys. Chem. Chem. Phys.* **2008**, *10*, 3335.
- Kamiya, M.; Hirata, S.; Valiev, M. *J. Chem. Phys.* **2008**, *128*, 074103.
- Dahlke, E. E.; Leverentz, H. R.; Truhlar, D. G. *J. Chem. Theory Comput.* **2008**, *4*, 33.
- Russ, N. J.; Crawford, T. D. *Phys. Chem. Chem. Phys.* **2008**, *10*, 3345.
- Li, W.; Li, S. *J. Chem. Phys.* **2004**, *121*, 6649.
- Li, W.; Piecuch, P.; Gour, J. R.; Li, S. *J. Chem. Phys.* **2009**, *131*, 114109.
- Saebø, S.; Pulay, P. *Annu. Rev. Phys. Chem.* **1993**, *44*, 213.
- Schütz, M.; Hetzer, G.; Werner, H. J. *J. Chem. Phys.* **1999**, *111* (13), 5691.
- Adler, T. B.; Werner, H.-J.; Manby, F. R. *J. Chem. Phys.* **2009**, *130*, 054106.
- Schütz, M. *J. Chem. Phys.* **2000**, *113* (22), 9986.
- Pisani, C.; Busso, M.; Capecchi, G.; Casassa, S.; Dovesi, R.; Maschio, L.; Zicovich-Wilson, C.; Schütz, M. *J. Chem. Phys.* **2005**, *122* (9).
- Maschio, L.; Usvyat, D.; Manby, F. R.; Casassa, S.; Pisani, C.; Schütz, M. *Phys. Rev. B* **2007**, *76* (7).
- Usvyat, D.; Maschio, L.; Manby, F. R.; Casassa, S.; Schütz, M.; Pisani, C. *Phys. Rev. B* **2007**, *76* (7).
- Subotnik, J. E.; Sodt, A.; Head-Gordon, M. *J. Chem. Phys.* **2006**, *125*, 074116.
- Subotnik, J. E.; Sodt, A.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *128*, 034103.
- Chwee, T. S.; Szilva, A. B.; Lindh, R.; Carter, E. A. *J. Chem. Phys.* **2008**, *128*, 224106.
- Fedorov, D. G.; Kitaura, K. *J. Chem. Phys.* **2005**, *123*, 134103.
- Li, W.; Piecuch, P.; Gour, J. R. *Theory and Applications of Computational Chemistry - 2008, AIP Conference Proceedings*, 2009; Vol. 1102, p 68.
- Hughes, T. F.; Flocke, N.; Bartlett, R. J. *J. Phys. Chem. A* **2008**, *112*, 5994.
- Stoll, H. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1992**, *46*, 6700.

- (37) Stoll, H. *J. Chem. Phys.* **1992**, *97*, 8449.
- (38) Nesbet, R. K. *Phys. Rev.* **1968**, *175*, 2.
- (39) Nesbet, R. K. *Adv. Chem. Phys.* **1969**, *14*, 1.
- (40) Friedrich, J.; Hanrath, M.; Dolg, M. *J. Phys. Chem. A* **2007**, *111*, 9830.
- (41) Friedrich, J.; Dolg, M. *J. Chem. Phys.* **2008**, *129*, 244105.
- (42) Friedrich, J.; Dolg, M. *J. Chem. Theory Comput.* **2009**, *5*, 287.
- (43) Yu, M.; Kalvoda, S.; Dolg, M. *Chem. Phys. Lett.* **1997**, *224*, 121.
- (44) Abdurahman, A.; Shukla, A.; Dolg, M. *J. Chem. Phys.* **2000**, *112*, 4801.
- (45) Stoll, H.; Paulus, B.; Fulde, P. *J. Chem. Phys.* **2005**, *123*, 144108.
- (46) Albrecht, M.; Paulus, B.; Stoll, H. *Phys. Rev. B* **1997**, *56*, 7339.
- (47) Doll, K.; Dolg, M.; Fulde, P.; Stoll, H. *Phys. Rev. B* **1995**, *52*, 4842.
- (48) Doll, K.; Dolg, M.; Fulde, P.; Stoll, H. *Phys. Rev. B* **1997**, *55*, 10282.
- (49) Kalvoda, S.; Paulus, B.; Dolg, M.; Stoll, H.; Werner, H.-J. *Phys. Chem. Chem. Phys.* **2001**, *3*, 514.
- (50) Paulus, B. *Int. J. Quantum Chem.* **2004**, *100*, 1026.
- (51) Friedrich, J.; Hanrath, M.; Dolg, M. *J. Chem. Phys.* **2007**, *126*, 154110.
- (52) Friedrich, J.; Hanrath, M.; Dolg, M. *Chem. Phys.* **2007**, *338*, 33.
- (53) Müller, C.; Herschend, B.; Hermansson, K.; Paulus, B. *J. Chem. Phys.* **2008**, *128*, 214701.
- (54) Müller, C.; Paulus, B.; Hermansson, K. *Surf. Sci.* **2009**, *603*, 2619.
- (55) Müller, C.; Hermansson, K.; Paulus, B. *Chem. Phys.* **2009**, *362*, 91.
- (56) Schmitt, I.; Fink, K.; Staemmler, V. *Phys. Chem. Chem. Phys.* **2009**, *11*, 11196.
- (57) Friedrich, J.; Hanrath, M.; Dolg, M. *Chem. Phys.* **2008**, *346*, 266.
- (58) Friedrich, J.; Hanrath, M.; Dolg, M. *J. Phys. Chem. A* **2008**, *112*, 8762.
- (59) Friedrich, J.; Coriani, S.; Helgaker, T.; Dolg, M. *J. Chem. Phys.* **2009**, *131*, 154102.
- (60) Friedrich, J.; Walczak, K.; Dolg, M. *Chem. Phys.* **2009**, *356*, 47.
- (61) Friedrich, J.; Tew, D.; Klopper, W.; Dolg, M. *J. Chem. Phys.* **2010**, *132*, 164114.
- (62) Foster, J. M.; Boys, S. F. *Rev. Mod. Phys.* **1960**, *32*, 300.
- (63) Angeli, C.; Del Re, G.; Persico, M. *Chem. Phys. Lett.* **1995**, *233*, 102.
- (64) Ahmadi, G. R.; Røgggen, I. *Theor. Chem. Acc.* **1997**, *97*, 41.
- (65) Alrichs, R.; Bär, M.; Baron, H.-P.; Bauernschmitt, R.; Böcker, S.; Ehrig, M.; Eichkorn, K.; Elliott, S.; Furche, F.; Haase, F.; Häser, M.; Horn, H.; Huber, C.; Huniar, U.; Kölmel, C.; Kollwitz, M.; Ochsenfeld, C.; Öhm, H.; Schäfer, A.; Schneider, U.; Treutler, O.; von Arnim, M.; Weigend, F.; Weis, P.; Weiss, H. *Turbomole 5.10*; Institut für Physikalische Chemie, Universität Karlsruhe: Karlsruhe, Germany, 2008.
- (66) Pipek, J.; Mezey, P. G. *J. Chem. Phys.* **1989**, *90*, 4916.
- (67) Karypis, G.; Kumar, V. *SIAM J. Sci. Comput.* **1998**, *20* (1), 359.
- (68) Friedrich, J.; Hanrath, M.; Dolg, M. *Z. Phys. Chem.* **2010**, in press.
- (69) Edmiston, C.; Ruedenberg, K. *Rev. Mod. Phys.* **1963**, *35*, 457.
- (70) Becke, A. D. *Phys. Rev. A* **1988**, *38* (6), 3098.
- (71) Perdew, J. P. *Phys. Rev. B* **1986**, *33* (12), 8822.
- (72) Treutler, O.; Ahlrichs, R. *J. Chem. Phys.* **1995**, *102*, 346.
- (73) Eichkorn, K.; Treutler, O.; Oehm, H.; Haeser, M.; Ahlrichs, R. *Chem. Phys. Lett.* **1995**, *240*, 283.
- (74) Eichkorn, K.; Weigend, F.; Treutler, O.; Ahlrichs, R. *Theor. Chem. Acc.* **1997**, *97*, 119.
- (75) Deglmann, P.; May, K.; Furche, F.; Ahlrichs, R. *Chem. Phys. Lett.* **2004**, *384*, 103.
- (76) Weigend, F.; Hättig, C. *J. Chem. Phys.* **2000**, *113*, 5154.
- (77) Weigend, F.; Häser, M.; Patzelt, H.; Ahlrichs, R. *Chem. Phys. Lett.* **1998**, *294*, 143.
- (78) Bulusu, S.; Yoo, S.; Apra, E.; Xantheas, S.; Zeng, X. C. *J. Phys. Chem. A* **2006**, *110*, 11781.
- (79) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007.
- (80) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796.

CT1000999

## A Subsystem TDDFT Approach for Solvent Screening Effects on Excitation Energy Transfer Couplings

Johannes Neugebauer,<sup>\*,†</sup> Carles Curutchet,<sup>‡,§</sup> Aurora Muñoz-Losa,<sup>||</sup> and Benedetta Mennucci<sup>||</sup>

*Gorlaeus Laboratories, Leiden Institute of Chemistry, Leiden University, P.O. Box 9502, 2300 RA Leiden, The Netherlands, Department of Chemistry, 80 St. George Street, Institute for Optical Sciences and Centre for Quantum Information and Quantum Control, University of Toronto, Toronto, Ontario M5S 3H6 Canada, Institut de Química Computacional and Departament de Química, Universitat de Girona, Campus Montilivi 17071 Girona, Catalonia, Spain, and Dipartimento di Chimica e Chimica Industriale, Università di Pisa, via Risorgimento 35, 56126 Pisa, Italy*

Received March 16, 2010

**Abstract:** We present a QM/QM approach for the calculation of solvent screening effects on excitation-energy transfer (EET) couplings. The method employs a subsystem time-dependent density-functional theory formalism [*J. Chem. Phys.* **2007**, *126*, 134116] and explicitly includes solvent excited states to account for the environmental response. It is investigated how the efficiency of these calculations can be enhanced in order to treat systems with very large solvation shells while fully including the environmental response. In particular, we introduce a criterion to select solvent excited states according to their approximate contribution weight to the environmental polarization. As a model system, we investigate the perylene diimide dimer in a water cluster in comparison to a recent polarizable QM/MM method for EET couplings in the condensed phase [*J. Chem. Theory Comput.* **2009**, *5*, 1838]. A good overall agreement in the description of the solvent screening is found. Deviations can be observed for the effect of the closest water molecules, whereas the screening introduced by outer solvation shells is very similar in both methods. Our results can thus be helpful to determine at which distance from a chromophore environmental response effects may safely be approximated by classical models.

### 1. Introduction

One of the fundamental steps in the primary events of photosynthesis is the transfer of excitation energy from a light-absorbing unit to a photosynthetic reaction center.<sup>1</sup> In the simplest case, this excitation-energy transfer (EET), which is a nonradiative process, involves the de-excitation of one chromophore (donor) together with the excitation of another pigment (acceptor).<sup>2</sup> The main mechanism for this transfer is a Coulomb interaction between the transition

densities of the two electronic transitions on the donor and acceptor, which for long distances can be described in terms of a transition-dipole interaction (Förster dipole coupling).<sup>3,4</sup> Other mechanisms can play a role in short separations of the donor and acceptor if there is considerable overlap of the monomer wave functions.<sup>5,6</sup> EET is an important effect not only in natural photosynthesis but also in artificial photosynthetic systems and optoelectronic devices.

While EET is a dynamic phenomenon, one of the essential ingredients in calculations of EET rate constants is the electronic coupling between the donor and acceptor transition, which is related to the energy difference between the coupled stationary electronic states (“excitonic states”) of the two chromophores. Consequently, much effort has been spent on the accurate calculation of transition densities of

\* Corresponding author e-mail: j.neugebauer@chem.leidenuniv.nl.

<sup>†</sup> Leiden University.

<sup>‡</sup> University of Toronto.

<sup>§</sup> Universitat de Girona.

<sup>||</sup> Università di Pisa.

pigment molecules involved in EET and their electronic couplings (“excitonic couplings”) during the past 10 to 15 years.<sup>7–22</sup> One of the open problems for a realistic description of excitation energy transfer rates is the inclusion of solvent effects, which is often just estimated on the basis of the dielectric constant of the environment. The screening of the Coulomb coupling by the solvent (or a general environment) can lead to considerable variations in the EET rates, especially for couplings at short and medium ranges.<sup>11,14,23–27</sup>

Recent investigations have addressed the possibilities of describing solvent screening effects on EET including more and more details of the environment. The approach presented in ref 14 is based on the polarizable continuum model (PCM)<sup>28</sup> and is able to consider the influence of the shape of the pigment molecules on the EET screening by the solvent. In contrast to this, the simple Förster approximation for the screening factor of EET rates considers a screening of point dipoles embedded in a dielectric continuum, which leads to problems at short range.

In ref 27, a polarizable QM/MM approach in combination with an ensemble averaging was developed for the simulation of solvent screening effects. It could be shown that, for homogeneous media, QM/MM and PCM results for the environmental screening are very similar. However, for heterogeneous media as present in proteins, QM/MM methods are expected to be more reliable, since they can model the environment in atomistic detail. While this is a clear advantage of the polarizable QM/MM approach, it requires a careful parametrization of the MM part. Furthermore, at very short range, the representation of the electrostatic effect of the environment in terms of point charges and induced dipoles as used in many QM/MM approaches may limit the overall accuracy of the calculation (cf. the benchmark study on protein effects on electronic spectra in ref 29). For certain specific effects, it may be necessary to include the relevant parts of the environment in the QM part of the calculation; examples addressing effects of hydrogen bonding, axial ligation, and effects of nearby charged residues on the absorption bands of bacteriochlorophyll molecules in a photosynthetic light-harvesting complex are given in refs 17 and 30. It should be noted that a direct assessment of such specific effects on the basis of experimental data is rather involved; for an example, see ref 31.

QM/MM approaches can be tested by comparing them to fully quantum chemical approaches. The study in ref 27 employed supermolecular quantum chemical calculations for this purpose, in which both interacting chromophores and the surrounding solvent were treated with configuration interaction singles (CIS). A very good agreement for both types of calculations was reported. However, supermolecular reference calculations are very demanding in terms of computer time and pose additional complications. In particular, for nonhybrid density functionals, many artificially low-lying charge-transfer excitations occur for solvated systems<sup>32</sup> due to the incorrect description of charge-transfer excitations when using exchange-correlation kernels obtained within the adiabatic local density approximation (ALDA) or the adiabatic generalized gradient approximation (AGGA).<sup>33–40</sup> This further increases the computational effort and hampers

the identification of the excitonic states, which is a prerequisite for extracting excitonic coupling constants from supermolecular calculations.

An alternative for the calculation of excitonic couplings from a purely quantum chemical approach is provided by subsystem methods within density functional theory. A subsystem formulation of density functional theory (DFT) was developed by Cortona in 1991<sup>41</sup> in the context of atomic subsystems in crystals (see also the earlier work by Senatore and Subbaswamy<sup>42</sup>). Subsequently, Wesolowski and Warshel proposed the so-called frozen-density embedding (FDE) method,<sup>43</sup> which can be regarded as a simplified and efficient subsystem method, in which only the electron density of an active part is optimized. The freeze-and-thaw method presented in ref 44 allows for a continuous transition between both extremes (fully frozen or fully relaxed environment) in the case of a partitioning into two subsystems, the active part and an environment. A fully relaxed subsystem DFT treatment involving many molecular subsystems was presented in ref 45, and a general setup allowing for different relaxation strategies of different subsystems is available<sup>46</sup> in the Amsterdam Density Functional program.<sup>47,48</sup>

An extension of the FDE method to excited states in terms of time-dependent DFT (TDDFT) was presented by Casida and Wesolowski.<sup>49</sup> On the basis of this work, a generalized subsystem TDDFT or coupled FDE-TDDFT (FDEc) approach was formulated by one of the present authors in refs 50 and 51, and efficient algorithms for calculating excitonic couplings for large aggregates of pigments were developed.<sup>17,50</sup> While the initial studies considered environmental effects only in terms of an effective embedding potential,<sup>17,50</sup> first applications that include solvent screening effects on exciton splittings of small models were presented in ref 52.

Here, we are going to investigate solvent screening effects in subsystem TDDFT in more detail and discuss possible strategies for the efficient solution of the computational problems involved. As a test system, we employ the intense low-lying  $\pi \rightarrow \pi^*$  transition of the solvated perylene diimide (PDI) dimer, which was investigated in ref 27.

## 2. Theory

The FDEc approach presented in refs 17 and 50 allows calculations of the excitation energies of a system composed of several molecules in two steps. First, the uncoupled FDE (FDEu) excitation energies are calculated; i.e., local excitations of all constituent molecules are obtained, embedded in an environment formed by all other molecules. These calculations employ the ground-state FDE embedding formalism as proposed by Wesolowski and Warshel,<sup>43</sup> in which the environmental density is obtained as a sum of all other molecules' densities. To obtain optimum subsystem densities, we iteratively apply freeze-and-thaw cycles<sup>44</sup> to all subsystems, so that effectively a variational subsystem density functional theory treatment is performed.<sup>41,45,53</sup> Local excited states are then obtained for all subsystems with the approximate form<sup>54</sup> of the FDE generalization to excited states<sup>49</sup> that restricts the response to the active subsystem only. In a second step, delocalized excited states of the entire



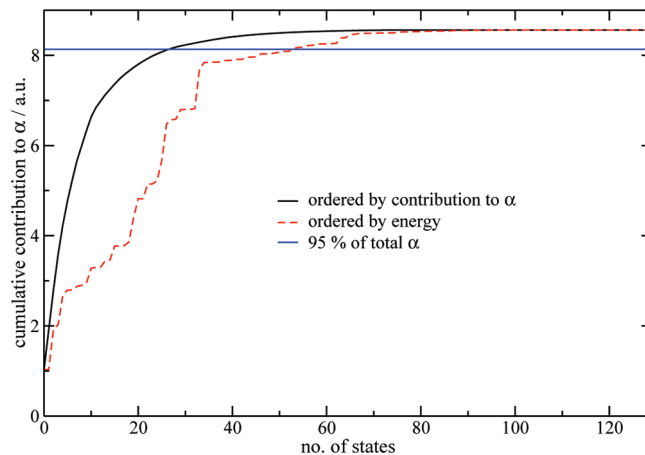
aggregate are calculated by coupling these local excitations following the subsystem TDDFT formalism presented in ref 17.

The main question that arises in approximate applications of the subsystem TDDFT formalism is the selection of states to be coupled. A direct connection to exciton coupling models can be made if only the relevant local excited states of the chromophores are coupled. These states will be called “reference states” in the following. In the present work, this would correspond to the intense low-lying  $\pi \rightarrow \pi^*$  excitation of the perylene diimide molecules. In ref 50, it was tested how the inclusion of additional excited states of the pigment molecules influences both the vertical excitation energies of the full system and the excitonic splitting between the reference states. A typical criterion that was used in the original implementation to determine states that have to be included<sup>50</sup> was the energy difference between a particular excited state and the reference states, since high-lying excited states only have a minor effect on the excitonic states. When modeling solvent effects, however, the situation is somewhat more complicated. The solvent screening can, in a linear-response TDDFT framework, be understood as a cumulative effect caused by many excited states of the solvent system. Consequently, a very large number of excited states would have to be included, which considerably increases the effort for FDEc calculations.

In order to achieve a reasonable representation of the solvent response, we have adopted the following strategy: We first determine how many excited states are necessary to represent the (isotropic) polarizability of a solvent molecule to a good accuracy in terms of the sum-over-states (SOS) expression (Hartree atomic units are used throughout)

$$\alpha(\omega) = \frac{2}{3} \sum_{\nu} \frac{\omega_{\nu}}{\omega_{\nu}^2 - \omega^2} |\mu_{0\nu}|^2 \quad (1)$$

where the sum runs over all excited states  $\nu$  with excitation energies  $\omega_{\nu}$  and transition dipole moments  $\mu_{0\nu}$  of the solvent molecule. Once these excitations are calculated for all (embedded) solvent molecules, they are sorted according to their contribution to the SOS polarizability expression in descending order. From this list of states, we choose the first  $k$  states, where the number  $k$  is determined in such a way that the cumulative contribution of these states is larger than a preselected threshold percentage  $p$  of the full SOS polarizability (obtained when including all precalculated excited states). Figure 1 shows the results for a water molecule (PBE/TZP) in the static limit. In this calculation, all singlet–singlet excitations within the TZP basis set have been calculated (130 in total). The dashed curve shows the cumulative polarizability contribution for the excited states when ordered by energy. This is a straightforward choice, since excited states are usually calculated with Davidson-type subspace iteration methods, which yield the lowest-energy transitions.<sup>55,56</sup> The solid line shows the results obtained if the excited states are ordered by their contribution to  $\alpha$ . About 90 states are needed to arrive at the converged isotropic polarizability value when the states are ordered by energy, whereas a similar convergence is already reached



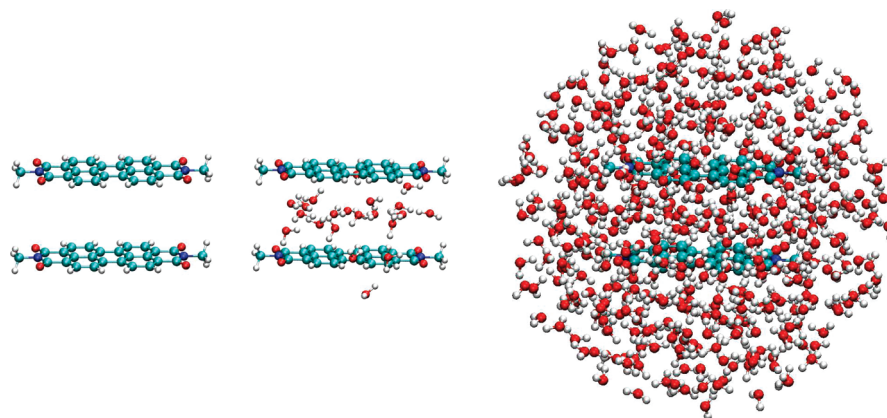
**Figure 1.** Cumulative contribution to the isotropic SOS polarizability of water (PBE/TZP, optimized structure, static limit) as a function of the number of excited states. Shown are results for contributions ordered according to the energy of the excited state (dashed line) and according to the contribution to the polarizability (solid line). The horizontal line represents a 95% threshold to the total polarizability.

with 60 states if the states are sorted by increasing contribution to  $\alpha$ . The horizontal line in Figure 1 indicates a 95% threshold of the polarizability. As will be shown below, this is typically sufficient for the calculation of screening effects on excitonic couplings. It can be seen that this threshold is reached with 53 states in the energy-sorted curve, whereas only 27 states are needed from the list ordered by polarizability contributions. This shows that a fairly small number of excited states may be sufficient to reproduce the polarizability of a water molecule. Nevertheless, the total number of coupled states increases tremendously when considering the solvent response. Another approximation that can be introduced is thus that solvent excited states are only coupled to the dye molecules' excited states, whereas intersolvent couplings are neglected. This approximation will be tested in section 4.

As discussed in refs 17 and 52, the elements of the matrix  $\tilde{\mathbf{Q}}$ , which describe the couplings between transitions on different subsystems, are calculated as

$$\tilde{\mathbf{Q}}_{\mu_A \nu_B} = \int d\mathbf{r}_1 \sum_{(ia)_A} 2F_{(ia)_A \mu_A} \sqrt{\omega_{(ai)_A}} \phi_{i_A}(\mathbf{r}_1) \delta v_{A, \nu_B}^{\text{ind}}(\mathbf{r}_1) \phi_{a_A}(\mathbf{r}_1) \quad (2)$$

Here,  $F_{(ia)_A \mu_A}$  is the solution factor describing the local excitation  $\mu_A$  in subsystem A,  $\phi_{i_A}$  and  $\phi_{a_A}$  are occupied and virtual, respectively, orbitals in subsystem A, and  $\omega_{(ai)_A}$  is their orbital energy difference;  $\delta v_{A, \nu_B}^{\text{ind}}$  is the potential that is induced in system A by the local electronic transition  $\nu_B$  of system B. The transition density of transition  $\nu_B$  enters the induced potential, while the sum over the orbital products  $F_{(ia)_A \mu_A}(\omega_{(ai)_A})^{1/2} \phi_{i_A} \phi_{a_A}$  can be identified with the transition density of transition  $\mu_A$ . In principle,  $\tilde{\mathbf{Q}}_{\mu_A \nu_B}$  should be symmetric, since only local response kernels are employed to calculate  $\delta v_{A, \nu_B}^{\text{ind}}$ . In practice, however, two different kinds of approximations are introduced that can make  $\tilde{\mathbf{Q}}$  nonsymmetric: The induced potential is constructed on the basis of a *fitted* transition density,<sup>50,57</sup> and the integration in eq 2 is performed in ADF by *numerical* integration.



**Figure 2.** Structure of the (solvated) PDI dimer from ref 27. Left, isolated dimer; middle, 7 Å solvation shell; right, 15 Å solvation shell.

The strategy introduced in refs 17 and 50 was to consider one of the subsystems (A) as the “active” subsystem and to evaluate its transition density exactly (but numerically). The other system is treated as the environmental system (B), and its transition density is fitted. In the current study, it is computationally advantageous to perform the numerical integration step for the smaller subsystems, i.e., the solvent molecules. We will refer to this as a “transpose construction” of  $\tilde{\Omega}$ .

In the FDEc calculations involving solvent response, no direct solvent contribution to the coupling can be calculated. The reason is that in the FDEc treatment the solvent response explicitly appears in the TDDFT eigenvalue problem in terms of solvent excited states, whereas it is contained implicitly in the interchromophore couplings in the QM/MMpol scheme and the PCM model. In order to extract excitonic coupling constants  $V$  from the FDEc calculations under the influence of the solvent response, we therefore use the expression derived from a secular determinant for Frenkel excitons of two-level chromophores,<sup>10</sup> as has been used in ref 27 to extract coupling constants from supermolecular calculations:

$$V = \frac{1}{2} \sqrt{(\omega_+ - \omega_-)^2 - (\omega_D - \omega_A)^2} \quad (3)$$

where  $\omega_{D,A}$  are the local excitation energies of the donor and acceptor, respectively, and  $\omega_{+,-}$  are the energies of the upper and lower, respectively, excitonic state.

### 3. Computational Details

All subsystem (TD)DFT calculations have been performed with a modified version<sup>17,50</sup> of the ADF 2008 program.<sup>47,48</sup> Supermolecular reference calculations and calculations on isolated molecules were carried out with the RESPONSE module of ADF.<sup>57</sup> We use the Perdew–Burke–Ernzerhof (PBE) exchange–correlation functional; for the LDA part, the Perdew–Wang (PW92) parametrization was employed, which corresponds to the default in the ADF 2009 version but is at variance with the ADF 2008 defaults. The TZP basis set from the ADF basis set library has been used for all ADF calculations. For the nonadditive kinetic energy contribution in subsystem DFT calculations, the so-called GGA97 generalized-gradient approximation (GGA) to the kinetic-energy

functional was employed.<sup>58</sup> It has the same functional form for the enhancement factor  $F(s)$  as the exchange functional of Perdew and Wang<sup>59</sup> and is therefore often denoted as PW91k. It was parametrized for the kinetic energy by Lembarki and Chermette.<sup>60</sup> For all (subsystem) TDDFT calculations, we applied the adiabatic local density approximation for the exchange–correlation kernel. In the case of subsystem TDDFT, also the kinetic-energy component of the kernel is approximated by the local-density (Thomas–Fermi) approximation.

QM/MM calculations with a polarizable force-field, denoted as QM/MMpol in the following, have been performed as described in ref 27 with a locally modified version of the Gaussian 03 package.<sup>61</sup> All QM/MMpol calculations employed the PBE exchange–correlation functional and a cc-pVTZ basis set.<sup>62</sup> The parametrization of the force-field part was adopted from ref 27 and consists of a set of distributed atomic polarizabilities calculated using the LoProp approach<sup>63</sup> combined with ESP charges fitted to the electrostatic potential, both obtained at the B3LYP/aug-cc-pVTZ level.

Test calculations on a PDI monomer resulted in excitation energies of 2.134 (PBE/TZP) and 2.145 eV (PBE/cc-pVTZ). The corresponding oscillator strengths are 0.548 (PBE/TZP) and 0.542 (PBE/cc-pVTZ). Excitonic splitting energies for the low-lying  $\pi \rightarrow \pi^*$  transitions based on supermolecular calculations were obtained as 0.0914 eV (PBE/TZP) and 0.0900 eV (PBE/cc-pVTZ). This shows that the results from the calculations with Slater-(TZP) and Gaussian-type (cc-pVTZ) basis sets are in very good agreement.

Graphics of the molecular structures were generated with the program VMD.<sup>64</sup>

### 4. Coupled Response of Solvated Dimers

The structure of the perylene diimide dimer investigated here is shown in Figure 2. It is the same structure investigated in ref 27 (intermolecular distance: 7 Å). Solvation shells with radii between 7 and 11 Å around the center of the pigment dimer have been considered in the initial tests to investigate the screening effect, while for the comparison with the QM/MMpol approach, extended solvation shells with up to 15 Å cutoffs (>1400 atoms in total) have been employed.

**Table 1.** Excitation Energy ( $E_u$ , in units of eV) and Oscillator Strength  $f_u$  of the Upper Excitonic State and Excitonic Splitting  $\Delta E$  (in units of  $\text{cm}^{-1}$ ) for the Isolated PDI Dimer with a Distance of 7 Å

	$E_u$	$\Delta E$	$f_u$
iso, uncoupled	2.134	0	0.548
FDEu	2.133	0	0.550
FDEc	2.178	734	1.100
super, iso	2.174	737	0.956

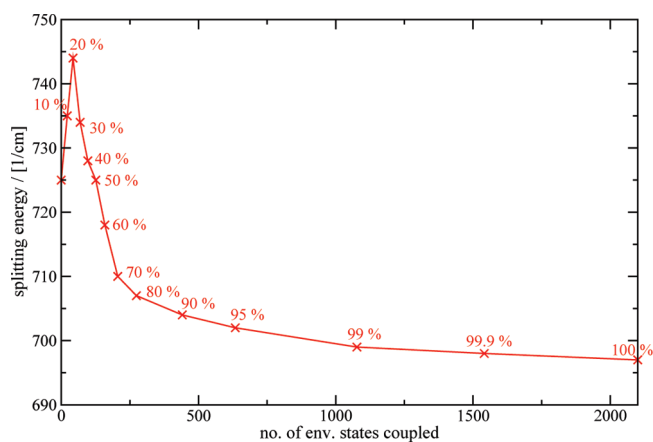
**Table 2.** Excitation Energy ( $E_u$ , in units of eV) and Oscillator Strength  $f_u$  of the Upper Excitonic State and Excitonic Splitting  $\Delta E$  (in units of  $\text{cm}^{-1}$ ) for the PDI Dimer for a 7 Å Solvation Shell with Different Thresholds  $\rho$  for the Cumulative Polarizability Contribution<sup>a</sup>

$\rho$	# states	$E_u$	$\Delta E$	$f_u$
FDEu		2.118	11	0.543
0.000	20	2.161	725	1.079
0.100	42	2.160	735	1.071
0.200	63	2.159	744	1.063
0.300	89	2.157	734	1.048
0.400	117	2.155	728	1.032
0.500	146	2.154	725	1.021
0.600	179	2.152	718	1.009
0.700	226	2.150	710	0.997
0.800	294	2.149	707	0.984
0.900	461	2.147	704	0.971
0.950	654	2.145	702	0.964
0.990	1097	2.144	699	0.959
0.999	1561	2.144	698	0.957
all	2120	2.144	697	0.957
super		2.115	644	0.768

<sup>a</sup> Also shown is the number of states that have to be coupled. FDEu denotes the uncoupled calculation; in the case of  $\rho = 0.000$ , only the 20 excited states calculated for the PDI molecules are coupled.

As a first test of the subsystem methodology to describe EET couplings, we calculated the excitation energies of the PDI dimer without the water environment in a supermolecular and a subsystem TDDFT calculation. The results are compared in Table 1. The FDEu data hardly differ from those of the isolated monomer calculations, as expected, because of the rather large distance of 7 Å between the monomers. The FDEc calculations lead to the expected splitting and reproduce the splitting energy of the supermolecular calculation very well (734 compared to 737  $\text{cm}^{-1}$ ). Also, the energy of the intense upper excitonic state from FDEc agrees nicely with the supermolecular results; the oscillator strength is, however, somewhat overestimated. A possible reason could be a basis set superposition effect in the supermolecular calculation (cf. the discussion of such effects in ref 51).

We now consider the smallest solvated system (7 Å solvation shell) to test the necessary approximations for the inclusion of the solvent response. In all calculations, 10 excited states per PDI molecule are included. Table 2 contains the results for different thresholds (percentages of the total SOS polarizability) for the cumulative polarizability contribution. The resulting splitting energies are also shown in Figure 3. The uncoupled FDE calculation leads to a small splitting between the two monomer transitions due to slightly different local environments in this snapshot. If the coupling between the PDI monomers is included in the calculation,



**Figure 3.** Splitting energies between the two excitonic states in the PDI dimer with a 7 Å solvation shell as a function of the threshold  $\rho$  for the cumulative polarizability contribution of the coupled states.

but no solvent response is taken into account, a splitting of 725  $\text{cm}^{-1}$  is observed. Including more and more environmental states first increases the energy gap to 744  $\text{cm}^{-1}$  if 20% of the polarizability is reproduced and then decreases it to a final value of 697  $\text{cm}^{-1}$  if all environmental states are included. If 95% of the polarizability is reproduced by the coupled states, the splitting energy is converged within 5  $\text{cm}^{-1}$  (0.7%). This threshold will be used in all subsequent calculations.

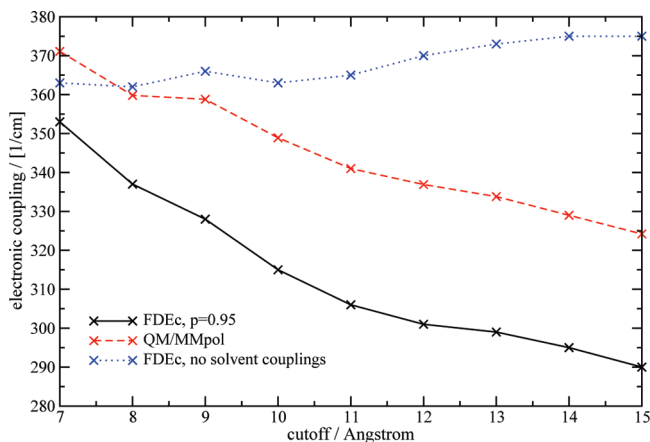
A supermolecular calculation on this system resulted in a somewhat smaller splitting energy of 644  $\text{cm}^{-1}$ . This discrepancy is not very large on an absolute scale. But there is a more pronounced difference in the results from FDEc and from the supermolecular calculation for the change in the splitting energy from the isolated PDI dimer to the solvated system studied here: In the FDEc case, the splitting reduces from 734 to 697  $\text{cm}^{-1}$ , whereas in the supermolecular case, it decreases from 737 to 644  $\text{cm}^{-1}$ . It should be noted, however, that the identification of the lower excitonic state in the supermolecular calculation is not entirely unambiguous. Many additional excited states appear, several of which are of the intermolecular charge-transfer type and thus significantly underestimated in our calculations (see the related problem in ref 32). Such difficulties do not appear in the absence of solvent molecules. Although an assignment can be made on the basis of the dominant orbital transitions, these orbital transitions also mix with other excitations for the PDI dimer, which are slightly lower in energy and would thus increase the splitting energy in the solvated case. Other reasons could be the monomer basis sets used to expand the (response) densities of the subsystems in the FDEc calculations. These basis sets are thus the same in the solvated and nonsolvated case, whereas the excitonic states in the supermolecular calculation benefit from the presence of basis functions at the solvent molecules. Also, the different accuracy of the available approximations for the kinetic-energy potential and kernel for different interaction strengths could play a role. These approximations will work better the smaller the subsystem density overlap is.<sup>65,66</sup> The PDI monomers in the isolated dimer are well separated, but water molecules appear quite close to the dye molecules



**Table 3.** Excitonic Coupling Constants (in units of  $\text{cm}^{-1}$ ) for the PDI Dimer for Different Sizes of the Solvation Shell (cutoffs in units of  $\text{\AA}$ )<sup>a</sup>

cutoff	nsc	nisc	$p = 0.95,$			QM/MMpol
			$p = 0.95$	transp	full	
7	363	359	351	353	348	371
8	362	345	335	337	332	360
9	366	335	328	328	322	359
10	363	319	315	315	310	349
11	365	307	306	306		341

<sup>a</sup> nsc, no solvent couplings; nisc, no inter-solvent couplings;  $p$  refers to the cumulative polarizability threshold; full, all solvent couplings fully included. The label “transp” refers to the transpose construction of  $\tilde{\Omega}$ .

**Figure 4.** Excitonic coupling constants for the PDI dimer with increasing solvation shell cutoffs. Results are shown for FDEc calculations neglecting any solvent couplings or employing a cumulative polarizability threshold of  $p = 0.95$  as well as for QM/MMpol calculations.

in the solvated dimer, so that the accuracy of the kinetic-energy contributions may be slightly worse in the solvated case.

As a next step, we investigate the dependence of the excitonic splitting on the size of the solvation shell. Table 3 contains the coupling constants calculated according to eq 3 for several different approximations. The column labeled “nsc” refers to calculations with “no solvent couplings”; i.e., the response of the environment is neglected. Only the environmental effect on the ground-state properties is taken into account in terms of the FDE potential (apart from a small modification of the exchange-correlation kernel, see ref 52 for details). As can be seen, there are only small effects on the calculated coupling constants ( $<5 \text{ cm}^{-1}$  compared to the coupling of  $367 \text{ cm}^{-1}$  found for the isolated PDI dimer), which means that there is hardly any effect of the environmental potential on the transition densities of the monomers. The additional data for cutoffs of 12 to 15  $\text{\AA}$  presented in Figure 4 indicate a slight increase, so that the coupling converges to  $375 \text{ cm}^{-1}$ , still within  $8 \text{ cm}^{-1}$  of the isolated dimer. In contrast to that, all other calculations predict a decrease in the coupling constants with increasing cutoff for the solvation shell. If only solute–solvent couplings are included in addition to the solute–solute couplings,  $V$  decreases by  $52 \text{ cm}^{-1}$  from  $359 \text{ cm}^{-1}$  ( $7 \text{\AA}$ ) to  $307 \text{ cm}^{-1}$

( $11 \text{\AA}$ ). If a cumulative polarizability threshold of  $p = 0.95$  is applied,  $V$  decreases by  $45 \text{ cm}^{-1}$  from  $351 \text{ cm}^{-1}$  to  $306 \text{ cm}^{-1}$ . This shows that the neglect of intersolvent couplings leads to the same qualitative behavior, although the deviation is larger for smaller solvation shells. As a reference, we also carried out fully coupled calculations, in which all solvent couplings are included. These calculations are still quite demanding and have therefore only been carried out for cutoffs up to  $10 \text{\AA}$ . The results show the same trend as the  $p = 0.95$  values but are systematically shifted by about  $3$  to  $6 \text{ cm}^{-1}$ .

Table 3 also shows the results obtained with a transpose construction of  $\tilde{\Omega}$  as defined in section 2. It can be seen that the deviation between the two different construction schemes is very small ( $0$  to  $2 \text{ cm}^{-1}$ ). Since the transpose construction is a great computational advantage for the system studied here, it was employed for the calculations presented in the following.

A comparison of the EET couplings calculated with FDEc ( $p = 0.95$ ) and the corresponding values obtained from the QM/MMpol calculations is presented in Figure 4. For cutoffs of  $10 \text{\AA}$  or larger, the splittings calculated with FDEc and QM/MMpol run more or less parallel, with an offset of about  $35 \text{ cm}^{-1}$ . For smaller cutoffs of the solvation shell, however, the difference between the two curves decreases to  $18 \text{ cm}^{-1}$  (cutoff of  $7 \text{\AA}$ ), and the two curves are not parallel anymore. In other words, the effect of outer solvation shells is described in the same way in FDEc and QM/MMpol calculations, whereas there is a slight quantitative disagreement for the nearest solvent molecules.

Interestingly, the QM/MMpol slight overestimation of the FDEc couplings is similar to that found in ref 27, where QM/MMpol was compared to full quantum chemical calculations. Also in this work, supermolecular calculations indicate smaller splittings. These findings suggest that short-range nonelectrostatic interactions between the dyes and the first solvation shell, neglected in the QM/MMpol scheme, seem to slightly attenuate the electronic coupling.

## 5. Solvent Screening Factors

In the QM/MMpol approach presented in ref 27, the EET couplings  $V$  are obtained as a sum of two terms:

$$V = V_s + V_{\text{explicit}} \quad (4)$$

where  $V_s$  is the Coulomb plus exchange-correlation interaction of the transition densities  $\rho_{D,A}^T$  of the solvated donor (D) and acceptor (A) systems:

$$V_s = \int d\mathbf{r} \int d\mathbf{r}' \rho_A^T(\mathbf{r}') \left( \frac{1}{|\mathbf{r} - \mathbf{r}'|} + f_{\text{xc}}(\mathbf{r}, \mathbf{r}') \right) \rho_D^T(\mathbf{r}) - \omega_0 \int d\mathbf{r} \rho_A^T(\mathbf{r}) \rho_D^T(\mathbf{r}) \quad (5)$$

The last term is an overlap contribution that arises because the interaction in ref 27 is treated as a perturbation of separated systems A and D; it is usually very small.<sup>14</sup> In the present study, it never contributes more than  $0.6 \text{ cm}^{-1}$  to the total coupling. The explicit solvent contribution  $V_{\text{explicit}}$  describes the interaction of systems A and D that is mediated by the environment. To be more precise, it is calculated as



**Table 4.** Excitonic Coupling Constants  $V_s$  (from QM/MMpol; in units of  $\text{cm}^{-1}$ ) for the PDI Dimer and Solvent Screening Factors  $s_{\text{QM/MMpol}}$  and  $s_{\text{FDEc}}$  as a Function of the Cutoff Distance for the Solvent Shell<sup>a</sup>

cutoff	$V_s$	$s_{\text{QM/MMpol}}$	$s_{\text{FDEc}}$
7	391	0.95	0.90
8	406	0.89	0.83
9	432	0.83	0.76
10	437	0.80	0.72
11	446	0.76	0.69
12	458	0.73	0.66
13	465	0.72	0.64
14	470	0.70	0.63
15	472	0.69	0.61

<sup>a</sup> Note that the same  $V_s$  values have been employed for both sets of solvent screening factors, since  $V_s$  is not directly available from FDEc calculations.

the Coulomb interaction of the transition density  $\rho_A^T$  with the dipoles induced in the environment by  $\rho_D^T$ ,<sup>27</sup>

$$V_{\text{explicit}} = - \sum_k \left( \int d\mathbf{r} \rho_A^T(\mathbf{r}) \frac{(\mathbf{r}_k - \mathbf{r})}{|\mathbf{r}_k - \mathbf{r}|^3} \right) \mu_k^{\text{ind}}(\rho_D^T) \quad (6)$$

The induced dipoles  $\mu_k^{\text{ind}}$  at positions  $\mathbf{r}_k$  are employed in the QM/MMpol model to simulate the polarization of the environment.

Solvent screening factors  $s$  can then be calculated as<sup>27</sup>

$$s = \frac{V}{V_s} = \frac{V_s + V_{\text{explicit}}}{V_s} \quad (7)$$

The main difference with respect to the FDEc approach is that the transition densities employed in the calculation of  $V_s$ , eq 5, are obtained for the solvated monomers *including* an environmental response contribution. The explicit contribution to the coupling thus reflects the differential solvent polarization when the interaction between the monomers is “switched on”.

In contrast to that, the FDEc couplings are calculated on the basis of local transition densities of the subsystems that neglect the environmental response contribution, and the *entire* solvent response enters the calculation of the total EET couplings. If we would calculate the solvent screening factor as the ratio between the FDEc results without solvent couplings and the fully coupled FDEc data, we would thus employ a different definition of the solvent screening factor (see below).

If we assume the  $V_s$  values from the QM/MMpol calculations, which we cannot directly access on the basis of the FDEc approach, and combine them with the total couplings  $V$  calculated from FDEc, we can determine the solvent screening factors. This is done in Table 4. The values obtained with the polarizability criterion  $p = 0.95$  have been employed for that purpose. Also shown are the QM/MMpol  $V_s$  values as well as the QM/MMpol solvent screening factors. In both cases, the solvent screening decreases, and the FDEc solvent screening factor is systematically lower than the QM/MMpol result, as could be expected from the coupling constants. Nevertheless, there is a fair agreement between the two sets of calculations, and the trend is clearly the same.

**Table 5.** Solvent Screening Factors  $\tilde{s}$  Calculated as the Ratio of the Excitonic Coupling in the Solvent Shell and the Coupling in a Vacuum for the PDI Dimer as a Function of the Cutoff Distance for the Solvent Shell

cutoff	$\tilde{s}_{\text{QM/MMpol}}$	$\tilde{s}_{\text{FDEc}}$
7	1.02	0.96
8	0.99	0.92
9	0.99	0.89
10	0.96	0.86
11	0.94	0.83
12	0.93	0.82
13	0.92	0.81
14	0.91	0.80
15	0.89	0.79

The above definition of the screening factor is consistent with the factor assumed in Förster theory, which scales a dipole–dipole interaction obtained from transition dipole moments measured for the noninteracting dyes in solution.<sup>67</sup> However, an alternative definition of the screening factor that accounts for the entire solvent effect, and thus allows a more in-depth comparison between the FDEc and QM/MMpol methods, is given by the ratio

$$\tilde{s} = \frac{V_{\text{solution}}}{V_{\text{vacuum}}} \quad (8)$$

of the coupling constant of the solvated dimer,  $V_{\text{solution}}$ , divided by the coupling constant of the dimer in a vacuum,  $V_{\text{vacuum}}$  ( $367 \text{ cm}^{-1}$  for FDEc and  $363 \text{ cm}^{-1}$  for QM/MMpol). Table 5 reports the results adopting this alternative solvent screening factor  $\tilde{s}$ . Interestingly, the coupling constant for a 7 Å cutoff with the QM/MMpol method is larger in solution than in the isolated dimer. The solvent screening factor  $\tilde{s}_{\text{QM/MMpol}}$  decreases by 12.6% from 1.02 (7 Å) to 0.89 (15 Å). The FDEc solvent screening factor  $\tilde{s}_{\text{FDEc}}$  shows a somewhat stronger decrease of 17.9% from 0.96 to 0.79. Both sets of screening factors are considerably larger than those obtained with the original definition shown in Table 4.

## 6. Conclusion

In this work, we have demonstrated that it is possible to include solvent screening effects into the calculation of excitonic splittings and EET couplings in the subsystem TDDFT formalism. Although the computational effort is considerably increased compared to calculations for isolated chromophores, several developments and approximations have been presented that allow an enhancement of the efficiency of the calculations. In particular, the number of solvent excited states needed to reproduce the environmental response effect could considerably be reduced by employing a polarizability-related criterion to select the coupled states. Furthermore, a transpose construction of the coupling matrix, in which the numerical integration step of the matrix elements is always carried out for the smaller subsystem, greatly reduces the computer time necessary for the calculation. Both procedures do not affect the magnitude of the calculated coupling constants significantly, and deviations were always within  $6 \text{ cm}^{-1}$  or 0.7 meV. This shows that the full response of environmental systems with more than 1000 atoms can

be treated accurately and fully quantum mechanically with the present approach. The results for the transpose construction also underline that the approximation of a symmetric coupling matrix made in ref 50 is well justified. Additional approximations can involve the neglect of intersolvent couplings, although this leads to slightly larger deviations from the fully coupled results.

FDEc and QM/MMpol agree rather well on the EET couplings of the solvated systems, and on their dependence on the size of the solvation shell. Discrepancies between the two approaches are on the order of 20 to 35 cm<sup>-1</sup> (6 to 10%). In particular, the effect of outer solvent molecules is very similar in both methods, whereas the deviations are a bit larger for smaller solvation shells.

Since FDEc and QM/MMpol describe the solvent response effects in different ways, it is not straightforward to calculate screening constants for FDEc in the way defined in ref 27. However, if the unscreened EET couplings are taken from the QM/MMpol calculation, then FDEc leads to a similar screening dependence on the cutoff radius of the solvation shell as QM/MMpol, although the predicted FDEc screening constants are somewhat smaller. This also holds for an alternative definition of the solvent screening as the ratio between the EET couplings in solution and in a vacuum, which in general leads to larger solvent screening constants.

This study thus indicates that there are small differences in the description of short-range electronic couplings between the subsystem TDDFT (FDEc) approach and the polarizable QM/MM approach. It also allows an estimation of the size of the solvation shell in which these differences become negligible. In the present example, it turned out that solvent molecules beyond the 10 Å cutoff have roughly the same effect in both approaches. Our work thus forms the basis for multiscale approaches to model the screening effect of general environmental systems on EET couplings that include the possibility to control the error introduced by different representations of different parts of the environment. This will be increasingly important in simulations of energy-transfer phenomena of protein–pigment complexes as occurring in natural photosynthetic systems.

**Acknowledgment.** J.N. is supported by a VIDI grant (700.59.422) of The Netherlands Organisation for Scientific Research (NWO) and acknowledges a computer time grant from the Stichting Nationale Computer Faciliteiten (NCF). C.C. acknowledges support from the Comissionat per a Universitats i Recerca of the Departament d'Innovació, Universitats i Empresa of the Generalitat de Catalunya, grant no. 2008BPB00108. A.M.-L. thanks support from the Spanish Ministerio de Ciencia e Innovación (Programa Nacional de Recursos Humanos del Plan Nacional I-D+I 2008–2011).

## References

- (1) Blankenship, R. E. *Molecular Mechanisms of Photosynthesis*; Blackwell Science: Oxford, 2002.
- (2) Scholes, G. D. *Annu. Rev. Phys. Chem.* **2003**, *54*, 57–87.
- (3) Förster, T. *Ann. Phys.* **1948**, *2*, 55.
- (4) Förster, T. Delocalized Excitation and Excitation Transfer. In *Modern Quantum Chemistry. Part III: Action of Light and Organic Crystals*; Sinanoğlu, O., Ed.; Academic Press: New York, 1965; pp 93–137.
- (5) Dexter, D. L. *J. Chem. Phys.* **1953**, *21*, 836–850.
- (6) Harcourt, R. D.; Scholes, G. D.; Ghiggino, K. P. *J. Chem. Phys.* **1994**, *101*, 10521–10525.
- (7) Krueger, B. P.; Scholes, G. D.; Fleming, G. R. *J. Phys. Chem. B* **1998**, *102*, 5378–5386.
- (8) Damjanović, A.; Ritz, T.; Schulten, K. *Phys. Rev. E* **1999**, *59*, 3293–3311.
- (9) Tretiak, S.; Middleton, C.; Chernyak, V.; Mukamel, S. *J. Phys. Chem. B* **2000**, *104*, 9540–9553.
- (10) Tretiak, S.; Middleton, C.; Chernyak, V.; Mukamel, S. *J. Phys. Chem. B* **2000**, *104*, 4519–4528.
- (11) Hsu, C.-P.; Fleming, G. R.; Head-Gordon, M.; Head-Gordon, T. *J. Chem. Phys.* **2001**, *114*, 3065–3072.
- (12) Wong, K. F.; Bagchi, B.; Rossky, P. J. *J. Phys. Chem. A* **2004**, *108*, 5752–5763.
- (13) Beenken, W. J. D.; Pullerits, T. *J. Chem. Phys.* **2004**, *120*, 2490–2495.
- (14) Iozzi, M. F.; Mennucci, B.; Tomasi, J.; Cammi, R. *J. Chem. Phys.* **2004**, *120*, 7029–7040.
- (15) Curutchet, C.; Mennucci, B. *J. Am. Chem. Soc.* **2005**, *127*, 16733–16744.
- (16) Madjet, M. E.; Abdurahman, A.; Renger, T. *J. Phys. Chem. B* **2006**, *110*, 17268–17281.
- (17) Neugebauer, J. *J. Phys. Chem. B* **2008**, *112*, 2207–2217.
- (18) Muñoz-Losa, A.; Curutchet, C.; Fdez. Galván, I.; Mennucci, B. *J. Chem. Phys.* **2008**, *129*, 034104.
- (19) Fink, R. F.; Pfister, J.; Zhao, H. M.; Engels, B. *Chem. Phys.* **2008**, *346*, 275–285.
- (20) Hsu, C.-P. *Acc. Chem. Res.* **2009**, *42*, 509–518.
- (21) Sagvolden, E.; Furche, F.; Köhn, A. *J. Chem. Theory Comput.* **2009**, *5*, 873–880.
- (22) Neugebauer, J. *ChemPhysChem* **2009**, *10*, 3148–3173.
- (23) Adolphs, J.; Renger, T. *Biophys. J.* **2006**, *91*, 2778–2797.
- (24) Scholes, G. D.; Curutchet, C.; Mennucci, B.; Cammi, R.; Tomasi, J. *J. Phys. Chem. B* **2007**, *111*, 6978–6982.
- (25) Russo, V.; Curutchet, C.; Mennucci, B. *J. Phys. Chem. B* **2007**, *111*, 853–863.
- (26) Curutchet, C.; Scholes, G. D.; Mennucci, B.; Cammi, R. *J. Phys. Chem. B* **2007**, *111*, 13253–13265.
- (27) Curutchet, C.; Muñoz-Losa, A.; Monti, S.; Kongsted, J.; Scholes, G. D.; Mennucci, B. *J. Chem. Theory Comput.* **2009**, *5*, 1838–1848.
- (28) Tomasi, J.; Mennucci, B.; Cammi, R. *Chem. Rev.* **2005**, *105*, 2999–3094.
- (29) Söderhjelm, P.; Husberg, C.; Strambi, A.; Olivucci, M.; Ryde, U. *J. Chem. Theory Comput.* **2009**, *5*, 649–658.
- (30) He, Z.; Sundström, V.; Pullerits, T. *J. Phys. Chem. B* **2002**, *106*, 11606–11612.
- (31) Timpmann, K.; Ellervee, A.; Pullerits, T.; Ruus, R.; Sundström, V.; Freiberg, A. *J. Phys. Chem. B* **2001**, *105*, 8436–8444.
- (32) Neugebauer, J.; Louwse, M. J.; Baerends, E. J.; Wesolowski, T. A. *J. Chem. Phys.* **2005**, *122*, 094115.

- (33) Dreuw, A.; Weisman, J. L.; Head-Gordon, M. *J. Chem. Phys.* **2003**, *119*, 2943–2946.
- (34) Tozer, D. *J. Chem. Phys.* **2003**, *119*, 12697–12699.
- (35) Tawada, Y.; Tsuneda, T.; Yanagisawa, S.; Yanai, T.; Hirao, K. *J. Chem. Phys.* **2004**, *120*, 8425–8433.
- (36) Gritsenko, O.; Baerends, E. J. *J. Chem. Phys.* **2004**, *121*, 655–660.
- (37) Maitra, N. T. *J. Chem. Phys.* **2005**, *122*, 234104.
- (38) Neugebauer, J.; Gritsenko, O.; Baerends, E. J. *J. Chem. Phys.* **2006**, *124*, 214102.
- (39) Ziegler, T.; Seth, M.; Krykunov, M.; Autschbach, J. *J. Chem. Phys.* **2008**, *129*, 184114.
- (40) Autschbach, J. *ChemPhysChem* **2009**, *10*, 1757–1760.
- (41) Cortona, P. *Phys. Rev. B* **1991**, *44*, 8454–8458.
- (42) Senatore, G.; Subbaswamy, K. R. *Phys. Rev. B* **1986**, *34*, 5754–5757.
- (43) Wesolowski, T. A.; Warshel, A. *J. Phys. Chem.* **1993**, *97*, 8050.
- (44) Wesolowski, T. A.; Weber, J. *Chem. Phys. Lett.* **1996**, *248*, 71–76.
- (45) Iannuzzi, M.; Kirchner, B.; Hutter, J. *Chem. Phys. Lett.* **2006**, *421*, 16–20.
- (46) Jacob, C. R.; Visscher, L. *J. Chem. Phys.* **2008**, *128*, 155102.
- (47) *Amsterdam Density Functional program*; Theoretical Chemistry, Vrije Universiteit: Amsterdam. URL: <http://www.scm.com> (access date: 01/17/2009).
- (48) te Velde, G.; Bickelhaupt, F. M.; Baerends, E. J.; van Gisbergen, S. J. A.; Fonseca Guerra, C.; Snijders, J. G.; Ziegler, T. *J. Comput. Chem.* **2001**, *22*, 931–967.
- (49) Casida, M. E.; Wesolowski, T. A. *Int. J. Quantum Chem.* **2004**, *96*, 577–588.
- (50) Neugebauer, J. *J. Chem. Phys.* **2007**, *126*, 134116.
- (51) Neugebauer, J. *J. Chem. Phys.* **2009**, *131*, 084104.
- (52) Neugebauer, J. *Phys. Reports* **2010**, *489*, 1–87.
- (53) Jacob, C. R.; Neugebauer, J.; Visscher, L. *J. Comput. Chem.* **2008**, *29*, 1011–1018.
- (54) Wesolowski, T. A. *J. Am. Chem. Soc.* **2004**, *126*, 11444–11445.
- (55) Davidson, E. R. *J. Comput. Phys.* **1975**, *17*, 87–94.
- (56) Murray, C. W.; Racine, S. C.; Davidson, E. R. *J. Comput. Phys.* **1992**, *103*, 382–389.
- (57) van Gisbergen, S. J. A.; Snijders, J. G.; Baerends, E. J. *Comput. Phys. Commun.* **1999**, *118*, 119–138.
- (58) Wesolowski, T. A. *J. Chem. Phys.* **1997**, *106*, 8516–8526.
- (59) Perdew, J. P. In *Electronic Structure of Solids*; Ziesche, P., Eschrig, H., Eds.; Akademie Verlag: Berlin, 1991; p 11.
- (60) Lembarki, A.; Chermette, H. *Phys. Rev. A* **1994**, *50*, 5328.
- (61) 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.
- (62) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007–1023.
- (63) Gagliardi, L.; Lindh, R.; Karlström, G. *J. Chem. Phys.* **2004**, *121*, 4494–4500.
- (64) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14.1*, 33–38.
- (65) Kiewisch, K.; Eickerling, G.; Reiher, M.; Neugebauer, J. *J. Chem. Phys.* **2008**, *128*, 044114.
- (66) Fux, S.; Kiewisch, K.; Jacob, C. R.; Neugebauer, J.; Reiher, M. *Chem. Phys. Lett.* **2008**, *461*, 353–359.
- (67) Knox, R. S.; van Amerongen, H. *J. Phys. Chem. B* **2002**, *106*, 5289–5293.

CT100138K

## Systematic Derivation of AMBER Force Field Parameters Applicable to Zinc-Containing Systems

Fu Lin and Renxiao Wang\*

*State Key Laboratory of Bioorganic Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai, People's Republic of China*

Received August 27, 2009

**Abstract:** Metal ions are indispensable for maintaining the structural stability and catalytic activity of metalloproteins. Molecular modeling studies of such proteins with force fields, however, are often hampered by the “missing parameter” problem. In this study, we have derived bond-stretching and angle-bending parameters applicable to zinc-containing systems which are compatible with the AMBER force field. A total of 18 model systems were used to mimic the common coordination configurations observed in the complexes formed by zinc-containing metalloproteins. The Hessian matrix of each model system computed at the B3LYP/6-311++G(2d,2p) level was then analyzed by Seminario's method to derive the desired force constants. These parameters were validated extensively in structural optimizations and molecular dynamics simulations of four selected model systems as well as one protein–ligand complex formed by carbonic anhydrase II. The best performance was achieved by a bonded model in combination with the atomic partial charges derived by the restrained electrostatic potential method. After some minor optimizations, this model was also able to reproduce the vibrational frequencies computed by quantum mechanics. This study provides a comprehensive set of force field parameters applicable to a variety of zinc-containing molecular systems. In principle, our approach can be applied to other molecular systems with missing force field parameters.

### 1. Introduction

Metalloproteins (or metalloproteinases) are a family of proteolytic enzymes whose catalytic mechanism involves a metal. The metal ion is an indispensable component for maintaining their enzymatic catalysis as well as structural stability.<sup>1</sup> Many metalloproteins are zinc-dependent. The zinc ion in such proteins often functions as a Lewis acid for the stabilization of reactants/intermediates or the occurrence of catalytic reactions. Although it is reported that zinc may adopt other types of coordination geometries,<sup>2,3</sup> it normally adopts a tetrahedral coordination geometry, in which the zinc ion is linked to the protein via three coordination bonds, while the fourth position is occupied by a labile water molecule or a bound ligand molecule. Matrix metalloproteinase (MMP), carbonic anhydrase (CA), alcohol dehydrogenase (AD), zinc-finger proteins are some well-studied zinc-

containing (Zn-containing) proteins. They play an essential role in the biosynthesis and metabolism of certain bioactive peptides and are relevant to a variety of critical diseases, including arthritis and cancer.<sup>4,5</sup> For example, MMPs conduct hydrolysis of the amide bonds on certain peptide substrates with the conserved zinc ion in the catalytic site. It has been demonstrated that MMPs regulate degradation of the extracellular matrix and control of angiogenesis, and thus selective inhibitors against MMPs may be used as promising anticancer therapies.<sup>6</sup>

Due to the important biological implications of Zn-containing proteins, molecular modeling is often employed to study the structures and functions of these proteins. Although today's computers are really powerful, modeling Zn-containing proteins in solvent with high-level quantum mechanics (QM) computations is still not quite possible. Thus, molecular mechanics (MM) is still the dominant approach for such tasks, although some combined QM/MM models<sup>7–10</sup> have been proposed as well. Unfortunately, most

\* Corresponding author. Telephone: 86-21-54925128. E-mail: wangrx@mail.sioc.ac.cn.



today's force fields do not always have appropriate parameters for metal atoms, which has become a practical obstacle for the molecular modeling studies of metalloproteins.

Researchers have developed various methods for tackling this "missing parameter" problem regarding metal atoms, in particular zinc. In general, there are three options: nonbonded, semibonded, and bonded models. The nonbonded model relies basically on electrostatic and van der Waals interactions instead of covalent bonds to maintain the coordination configuration of zinc ion during simulation.<sup>11</sup> It is straightforward to incorporate a nonbonded model into an established force field. However, such a model could be sensitive to the choice of atomic charge models. Due to the long-range nature of electrostatic forces, the zinc ion tends to get close to any negatively charged amino acid residues. Consequently, the zinc ion may have problems in retaining a low coordination number or even escapes from the coordination center.<sup>12</sup> Another problem with this type of models is that they cannot take account for charge transfer and polarization effects very well.<sup>13</sup>

Semibonded models were originally proposed by Pang et al.,<sup>14,15</sup> which are interesting patches to the nonbonded models. A semibonded model places four dummy atoms around the zinc ion, which are covalently connected to the zinc ion in a tetrahedral geometry. The zinc ion is assigned only van der Waals parameters, and its +2e charge is evenly distributed among the four dummy atoms. Interactions between the dummy atoms and amino acid residues are then computed using the conventional electrostatic interaction term. Semibonded models are also relatively convenient to be incorporated into an established force field. Compared to the nonbonded models, they are more suitable for modeling the tetrahedral coordination configuration of a Zn-containing system. However, they basically share the same shortcomings as nonbonded models; they are sensitive to the choice of atomic charge models and cannot be applied to other tasks, such as normal-mode analysis. How to properly set the parameters for the connections between zinc and dummy atoms is another matter of concern.

The bonded models<sup>16</sup> treat the connections between zinc and its ligands as covalent bonds. An obvious advantage of such models is that they can preserve the tetrahedral coordination configuration of zinc even in long-time simulations. If necessary, they can also be applied to other possible coordination configurations of zinc. A disadvantage of bonded models is that it is not convenient to use them to simulate the interconversions between different coordination configurations, since connection tables are kept fixed during simulation. Nevertheless, this is normally not a concern for molecular modeling studies of Zn-containing proteins. So far some researchers have derived force field parameters applicable to Zn-containing systems for bonded models through various approaches.<sup>17–24</sup> A common practice is to perform frequency analyses on Zn-containing model systems by high-level quantum mechanics computations, and then the diagonal elements in the resulting Hessian matrix are taken as the desired force constants of the bonds or the angles in which zinc is participated. This approach requires the Hessian matrix to be given in internal coordinates, which is

relatively complicated. Besides, a problem observed with this approach is that different settings of internal coordinates may lead to different results.<sup>25</sup>

Seminario et al. proposed an alternative method<sup>25–29</sup> by which a force constant may be derived from a Hessian matrix based on the Cartesian coordinates. This method retrieves the  $3 \times 3$  submatrices relevant to the atom pairs of interests from a given Hessian matrix. Then, the force constant for any internal coordinate (bond stretching, angle bending, and torsional angle) can be derived from the eigenvalues and eigenvectors of these submatrices after some mathematical transformations. Results produced by this method are obviously independent of the choice of internal coordinates. In their original study,<sup>25</sup> Seminario et al. has applied this method to some very simple molecules, such as water and nitrogen dioxide. Later, Ryde et al.<sup>29</sup> employed this method to derive force field parameters which were used for the refinement of crystal structures of metal-containing enzymes, such as ferrochelatase and iron superoxide dismutase. Bautista et al.<sup>26</sup> employed this method to derive force field parameters applicable to polyalanine peptides. Collectively, these studies have demonstrated the value of Seminario's method. Expanding the application of this method to more challenging problems is certainly intriguing.

In our study, we adopted Seminario's method to derive force field parameters applicable to various Zn-containing molecular complexes. These parameters, including equilibrium bond lengths and bond angles as well as force constants of bond stretching and angle bending, are compatible with the popular AMBER force field.<sup>30</sup> These parameters were validated on four simple Zn-containing model systems. Our results indicate that application of these parameters to the AMBER force field well reproduced the three-dimensional structures and vibrational frequencies of these model systems. These parameter were further applied to the molecular dynamics simulations on one complex structure formed by carbonic anhydrase II and produced encouraging results. We expect that the parameters derived in our study are applicable to the refinement of crystal and NMR structures of Zn-containing metalloproteins and the molecular modeling studies on the binding of such proteins to their ligand molecules. Compared to other researchers' previous studies, our study provides a more comprehensive set of force field parameters for a variety of Zn-containing complexes so that they can be readily applied without additional adjustment or optimization. More importantly, our study has demonstrated a complete approach to the deduction and validation of force field parameters with Seminario's method. The application of this approach is certainly not limited to Zn-containing molecular systems.

## 2. Methods

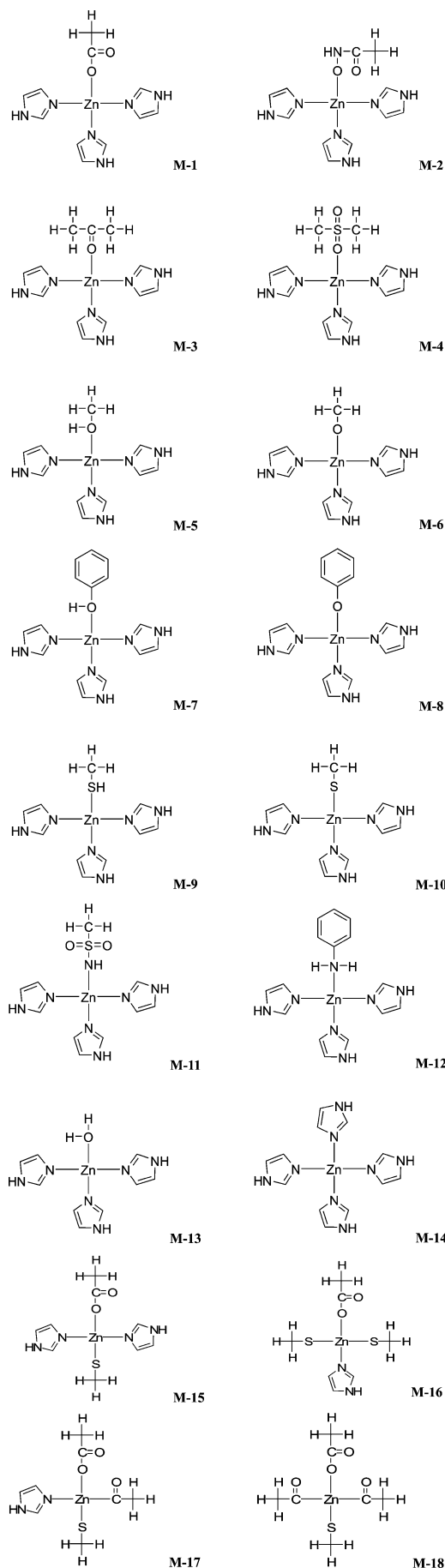
In our study, a total of 18 molecular complexes were used to mimic the typical situations in the binding of Zn-containing proteins with their ligand molecules. Force field parameters relevant to zinc were derived from the outcomes of QM computations on these model systems with Seminario's method. The derived parameters were then evaluated on four model systems to see if they could reproduce the

three-dimensional structures and vibrational frequencies of these model systems. They were also evaluated in molecular dynamics (MD) simulations of one complex structure formed by carbonic anhydrase II, a Zn-containing metalloprotein. QM computations were performed by using the Gaussian 03 software.<sup>31</sup> All of the other major computations, including energy minimization, MD simulation, and frequency analysis described in the following sections, were performed by using the AMBER software (version 9)<sup>32</sup> on a Linux cluster based on Intel Xeon 5345 processors.

**2.1. Selection of Zinc-Containing Model Systems and QM Computations.** The entire Protein Data Bank (PDB)<sup>33</sup> released by January 1, 2009, consisting of about 55 000 structures, was screened with an in-house computer program to retrieve the Zn-containing metalloproteins of our interests. Only the metalloproteins containing one zinc ion inside the binding pocket and one bound small-molecule ligand, i.e. Zn-containing protein–ligand complexes, were considered during this process. In addition, the zinc ion must be in contact with at least four nonhydrogen atoms within a distance cutoff of 2.8 Å, among which at least one had to be on the ligand molecule. Only crystal structures with overall resolution equal to or better than 2.5 Å were considered in order to impose a control on the quality of these complex structures. The total number of the Zn-containing protein–ligand complexes meeting the above criteria was 1004.

A survey on these complex structures revealed that in most cases the zinc ion was bound with three His residues on the protein side, although Cys, Glu, and Asp residues were observed in some cases. For the sake of convenience, we used three imidazole molecules to mimic the side chains of three His residues on the protein side. On the ligand side, a variety of chemical moieties were found in direct bonding with the zinc ion. Accordingly, a total of 12 small molecules (M-1 to M-12 in Figure 1) were used as models in our study, which covered the majority of such moieties identified in our survey. Six additional model systems (M-13 to M-18 in Figure 1) were considered to represent other mixed coordination centers, in which an acetic acid molecule was used to mimic the side chain of an Asp/Glu residue and a methanethiol molecule was used to mimic the side chain of a Cys residue. In fact, M-1, M-9, and M-10 also can be considered as mixed coordination centers. Therefore, the outcomes of our study can be applied to the modeling of a wider range of Zn-containing proteins.

QM computations were then performed on the 18 model systems summarized in Figure 1. These model systems were optimized by using the B3LYP method,<sup>34–38</sup> a popular density functional theory (DFT) method, with the 3-21G, 6-31+G(d, p), and then 6-311++G(2d, 2p) basis sets in a stepwise manner. Finally, frequency analysis was performed at the B3LYP/6-311++G (2d, 2p) level to confirm that the optimized structure was a true energy minimum without any imaginary frequency. The frequency analysis also produced the necessary raw data required by the following step.



**Figure 1.** Eighteen Zn-containing model systems considered in this study.

**2.2. Derivation of Zinc-Related Force Field Parameters.** We aimed at deriving parameters for zinc that are compatible with the AMBER force field. The potential energy function used by AMBER is

$$E = \sum_{\text{bonds}} K_r(r - r_{\text{eq}})^2 + \sum_{\text{angles}} K_\theta(\theta - \theta_{\text{eq}})^2 + \sum_{\text{torsions}} \frac{V_n}{2}(1 + \cos[n\phi - \gamma]) + \sum_{i < j}^{\text{atoms}} \left( \frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^6} \right) + \sum_{i < j}^{\text{atoms}} \frac{q_i q_j}{\epsilon R_{ij}} \quad (1)$$

The five terms in the above equation compute the energies of bond stretching, angle bending, torsion angles and non-bonded van der Waals and electrostatic interactions, respectively. Detailed explanations on the parameters in the above equation can be found elsewhere.<sup>30</sup> Note that in the current AMBER force field implemented in the AMBER software package (version 9)<sup>32</sup> only the van der Waals parameters of nonbonded model for zinc are provided. In this study, we derived bond-stretching and angle-bending parameters of bonded model for zinc. The equilibrium length of each bond and the equilibrium value of each bond angle in which zinc participated were obtained directly from the three-dimensional structures of the 18 model systems listed in Figure 1, which were fully optimized with extensive QM computations. The force constants of bond stretching and angle bending were derived with the method proposed by Seminario et al. A brief description of this method is given below for the sake of readers. More details can be found in the original reference.<sup>25</sup>

In order to derive the force constants of bond stretching and angle bending required in eq 1, a Hessian matrix in Cartesian coordinates was extracted from the outcomes of frequency analysis of each model system:

$$[\delta F] = -[k] \times [\delta x] \quad (2)$$

Here,  $[k]$  denotes for the Hessian matrix of a system composed of  $N$  atoms,  $[\delta x]$  denotes for the vector of the displacements in Cartesian coordinates, and  $[\delta F]$  denotes for the vector of resulting reaction forces. The full form of eq 2 is

$$\begin{bmatrix} \delta F_1 \\ \delta F_2 \\ \delta F_3 \\ \vdots \\ \delta F_{3N} \end{bmatrix} = - \begin{bmatrix} \frac{\partial^2 E}{\partial x_1^2} & \frac{\partial^2 E}{\partial x_1 \partial x_2} & \frac{\partial^2 E}{\partial x_1 \partial x_3} & \dots & \frac{\partial^2 E}{\partial x_1 \partial x_{3N}} \\ \frac{\partial^2 E}{\partial x_2 \partial x_1} & \frac{\partial^2 E}{\partial x_2^2} & \frac{\partial^2 E}{\partial x_2 \partial x_3} & \dots & \frac{\partial^2 E}{\partial x_2 \partial x_{3N}} \\ \frac{\partial^2 E}{\partial x_3 \partial x_1} & \frac{\partial^2 E}{\partial x_3 \partial x_2} & \frac{\partial^2 E}{\partial x_3^2} & \dots & \frac{\partial^2 E}{\partial x_3 \partial x_{3N}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 E}{\partial x_{3N} \partial x_1} & \frac{\partial^2 E}{\partial x_{3N} \partial x_2} & \frac{\partial^2 E}{\partial x_{3N} \partial x_3} & \dots & \frac{\partial^2 E}{\partial x_{3N}^2} \end{bmatrix} \times \begin{bmatrix} \delta x_1 \\ \delta x_2 \\ \delta x_3 \\ \vdots \\ \delta x_{3N} \end{bmatrix} \quad (3)$$

According to the Seminario's method, the bond-stretching force constant of bond A–B can be derived from the Hessian matrix as a  $3 \times 3$  matrix, i.e.  $[k_{AB}]$ :

$$[\delta F_A] = -[k_{AB}] \times [\delta x_B] \quad (4)$$

$$\begin{bmatrix} \delta F_{Ax} \\ \delta F_{Ay} \\ \delta F_{Az} \end{bmatrix} = - \begin{bmatrix} \frac{\partial^2 E}{\partial x_A \partial x_B} & \frac{\partial^2 E}{\partial x_A \partial y_B} & \frac{\partial^2 E}{\partial x_A \partial z_B} \\ \frac{\partial^2 E}{\partial y_A \partial x_B} & \frac{\partial^2 E}{\partial y_A \partial y_B} & \frac{\partial^2 E}{\partial y_A \partial z_B} \\ \frac{\partial^2 E}{\partial z_A \partial x_B} & \frac{\partial^2 E}{\partial z_A \partial y_B} & \frac{\partial^2 E}{\partial z_A \partial z_B} \end{bmatrix} \times \begin{bmatrix} \delta x_B \\ \delta y_B \\ \delta z_B \end{bmatrix} \quad (5)$$

The differential of force in eq 5, i.e.,  $[\delta F]$ , represents the responding force on atom A due to a displacement in the coordinates of atom B. Diagonalization of the  $[k_{AB}]$  matrix gives the eigenvalues  $\lambda_i^{AB}$  and the corresponding eigenvectors  $v_i^{AB}$ .

$$k_{AB} = \sum_{i=1}^3 \lambda_i^{AB} |u^{AB} \cdot v_i^{AB}| \quad (6)$$

Here,  $k_{AB}$  is the harmonic bond stretching force constant for bond A–B;  $u^{AB}$  is the normalized vector pointing from atoms A to B. It should be noted that  $k_{AB} = 2K_r$  ( $K_r$  is the force constant of bond stretching used in eq 1).

Similarly, the angle-bending force constant  $k_\theta$  for angle  $\angle ABC$  can be derived by considering the responding forces on atoms A and C due to a displacement in the coordinates of atom B:

$$\begin{bmatrix} \delta F_{Ax} \\ \delta F_{Ay} \\ \delta F_{Az} \end{bmatrix} = - \begin{bmatrix} \frac{\partial^2 E}{\partial x_A \partial x_B} & \frac{\partial^2 E}{\partial x_A \partial y_B} & \frac{\partial^2 E}{\partial x_A \partial z_B} \\ \frac{\partial^2 E}{\partial y_A \partial x_B} & \frac{\partial^2 E}{\partial y_A \partial y_B} & \frac{\partial^2 E}{\partial y_A \partial z_B} \\ \frac{\partial^2 E}{\partial z_A \partial x_B} & \frac{\partial^2 E}{\partial z_A \partial y_B} & \frac{\partial^2 E}{\partial z_A \partial z_B} \end{bmatrix} \times \begin{bmatrix} \delta x_B \\ \delta y_B \\ \delta z_B \end{bmatrix} \quad (7)$$

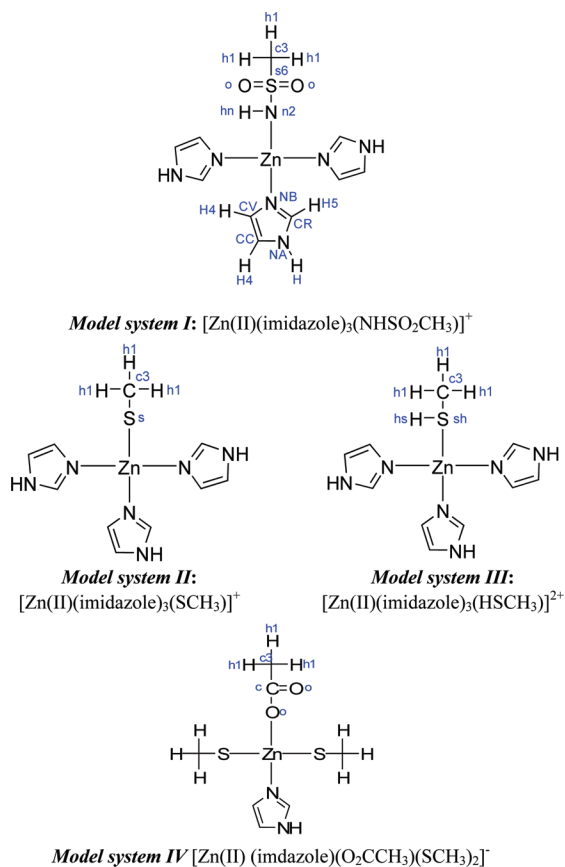
$$\begin{bmatrix} \delta F_{Cx} \\ \delta F_{Cy} \\ \delta F_{Cz} \end{bmatrix} = - \begin{bmatrix} \frac{\partial^2 E}{\partial x_C \partial x_B} & \frac{\partial^2 E}{\partial x_C \partial y_B} & \frac{\partial^2 E}{\partial x_C \partial z_B} \\ \frac{\partial^2 E}{\partial y_C \partial x_B} & \frac{\partial^2 E}{\partial y_C \partial y_B} & \frac{\partial^2 E}{\partial y_C \partial z_B} \\ \frac{\partial^2 E}{\partial z_C \partial x_B} & \frac{\partial^2 E}{\partial z_C \partial y_B} & \frac{\partial^2 E}{\partial z_C \partial z_B} \end{bmatrix} \times \begin{bmatrix} \delta x_B \\ \delta y_B \\ \delta z_B \end{bmatrix} \quad (8)$$

Then, the angle-bending force constant can be derived as:

$$\frac{1}{k_\theta} = \frac{1}{d_{AB}^2 \sum_{i=1}^3 \lambda_i^{AB} |u^{PA} \cdot v_i^{AB}|} + \frac{1}{d_{CB}^2 \sum_{i=1}^3 \lambda_i^{CB} |u^{PC} \cdot v_i^{CB}|} \quad (9)$$

Here,  $d_{AB}$  and  $d_{CB}$  are the distances between atoms A–B and C–B, respectively;  $u^{PA} = u_N \times u^{AB}$ ;  $u^{PC} = u^{CB} \times u_N$ ;  $u_N = (u^{CB} \times u^{AB}) / (|u^{CB} \times u^{AB}|)$ ;  $u^{AB}$ ,  $u^{CB}$ ,  $\lambda_i^{AB}$ ,  $\lambda_i^{CB}$ ,  $v_i^{AB}$ , and  $v_i^{CB}$  have similar meanings as in eq 6. Note that  $k_\theta = 2K_\theta$  ( $K_\theta$  is the angle-bending force constant in eq 1).

All of the 18 model systems summarized in Figure 1 were processed, as described above, using in-house computer programs. Note that, in principle, the harmonic force



**Figure 2.** Four model systems used for the validation of force field parameters. The AMBER atom type of each atom is labeled in lower cases.

constants of torsion angles can also be obtained through a similar procedure. However, the harmonic force constants are not compatible with the Fourier form of the torsion energy term in the current AMBER force field. Thus, we did not attempt to derive parameters for torsion angles which include zinc as one of the four component atoms. Neglecting the contributions of such torsion angles is actually not a major problem. After all, the bonded model was adopted in our study for modeling Zn-containing systems. Due to the symmetric tetrahedral geometry of the zinc coordination center, chemical moieties in bonding with the zinc ion are quite rigid and usually devoid of significant torsional freedom.<sup>20</sup>

**2.3. Validation of the Derived Parameters on Model Systems.** *Selection of the Model Systems and Force Field Models.* Four model systems, which were referred to as *model systems I–IV* (Figure 2), were selected out of the 18 typical Zn-containing model systems for validating the force field parameters derived in our study. The ligand molecule in *model system I* consists of a sulfonamide moiety, which is normally believed to be deprotonated upon coordination with zinc.<sup>37</sup> This moiety is observed in many ligands bound to Zn-containing proteins. For example, the core part of the complex formed between carbonic anhydrase II and 5-dimethylamino-naphtha-lene-1- sulfonamide (PDB entry 1OKL), which was used as an example later in our study for validation purposes, has exactly the same chemicals structure as *model system I*. The ligand molecules in both *model*

*systems II and III* consist of a thiol moiety, another common moiety for forming coordination bonds with zinc. This moiety also mimics the side chain of a Cys residue. The difference between *model systems II and III* was that the thiol moiety in *system II* was sketched in the deprotonated form; while the counterpart in *system III* was sketched in the neutral form. *Model system IV* presents a mixed coordination center, in which the three residues in bonding with zinc includes one His and two Cys residues. The ligand molecule in this model system is an acetic acid molecule in the deprotonated form. It can be considered as an Asp or Glu residue as well. In addition, *model system IV* bears an overall negative charge, unlike the other three model systems (Figure 2).

The performance of five force field models, including both bonded and nonbonded models (Table 1), were then evaluated on all four model systems. For the three bonded models (FF-1, FF-2, and FF-5), the force field parameters for zinc derived in our study (Table 2) were applied. For the nonbonded models (FF-3 and FF-4), parameters for the residues on the protein side were taken from the AMBER FF03 parameter set,<sup>39</sup> while parameters for the ligand molecule ( $^-\text{NHSO}_2\text{CH}_3$ ,  $^-\text{SCH}_3$ ,  $\text{HSCH}_3$ , and  $^-\text{O}_2\text{CCH}_3$ ) were taken from the AMBER GAFF parameter set.<sup>40,41</sup> The van der Waals parameters for zinc in all five force field models were set as:  $\sigma = 1.10 \text{ \AA}$  and  $\epsilon = 0.0125 \text{ kcal/mol}$ , which were cited from Merz's study.<sup>42</sup>

Since the choice of appropriate atomic charges, which are required to compute electrostatic energies, is another common argument in modeling Zn-containing systems, both the bonded and nonbonded models were tested in combination with two atomic partial charge schemes (Figure 3). The first scheme employed the RESP method<sup>43</sup> to derive atomic partial charges on the entire molecular system, including zinc, from the outcomes of QM computations at the B3LYP/6-311++G(2d,2p) level. This task was conducted with the RESP fitting protocol implemented in the AMBER program. Note that all QM computations in our study were conducted at the B3LYP/6-311++G(2d,2p) level. Thus, we did not repeat our computations at the HF/6-31G(d) level, which are typically supplied to the RESP method as inputs, to avoid possible inconsistency at other aspects. This charge scheme will be referred to as the “*RESP charges*” throughout this article. In the second scheme, zinc was assigned a formal charge of +2e. The atomic charges on the bonding residues were taken from the “*template charges*” for His and Cys residues in the AMBER FF03 parameter set. For the small-molecule ligand, template atomic charges are not available in the AMBER force field. Thus, atomic charges on the ligand molecule ( $^-\text{NHSO}_2\text{CH}_3$ ,  $^-\text{SCH}_3$ ,  $\text{HSCH}_3$ , and  $^-\text{O}_2\text{CCH}_3$ ) were also derived from the outcomes of QM computations at the B3LYP/6-311++G(2d,2p) level by using the RESP method. This scheme is in fact the standard practice employed by most common users of the AMBER program in the molecular modeling studies of metal-containing systems, and it will be referred to as the “*formal charges*” throughout this article.

*Structural Optimizations on Model Systems.* The five force field models were applied to all four model systems first to test how well they could reproduce the structures of these



**Table 1.** Reproduction of the Structures of Four Zn-Containing Model Systems by Five Different Force Field Models

force field model			rmsd (Å) <sup>a</sup>											
			model system I			model system II			model system III			model system IV		
symbol	bonding model	atomic charges <sup>b</sup>	in vacuum <sup>c</sup>	in vacuum <sup>d</sup>	in water <sup>d</sup>	in vacuum <sup>c</sup>	in vacuum <sup>d</sup>	in water <sup>d</sup>	in vacuum <sup>c</sup>	in vacuum <sup>d</sup>	in water <sup>d</sup>	in vacuum <sup>c</sup>	in vacuum <sup>d</sup>	in water <sup>d</sup>
FF-1	bonded	RESP	0.053	0.053	0.054	0.107	0.107	0.090	0.079	0.076	0.071	0.216	0.205	0.159
FF-2	bonded	formal	0.656	0.604	0.554	0.141	0.141	0.143	0.096	0.096	0.535	0.396	0.383	0.356
FF-3	nonbonded	RESP	1.001	0.293	0.275	0.091	0.085	0.104	16.347	16.424	0.236	1.795	1.778	0.245
FF-4	nonbonded	formal	0.255	0.255	0.240	0.266	0.265	0.267	0.231	0.231	0.237	0.358	0.336	0.287
FF-5 <sup>e</sup>	bonded	RESP	0.054	0.053	0.056	0.107	0.107	0.090	0.083	0.076	0.071	0.218	0.207	0.161

<sup>a</sup> Rmsd values were computed by considering the coordinates of zinc and the four atoms in direct bonding with zinc. The structure optimized at the B3LYP/6-311++G(2d, 2p) level was used as the reference. <sup>b</sup> See Figure 4. <sup>c</sup> Structural optimization started from an arbitrary structure. <sup>d</sup> Structural optimization started from a structure preoptimized at the B3LYP/6-311++G(2d, 2p) level. <sup>e</sup> A variant based on FF-1 which was specially optimized to better reproduce vibrational frequencies.

model systems preoptimized at the B3LYP/6-311++G(2d,2p) level. On each model system, the structural optimization was first started from the preoptimized structure by QM. To further test the robustness of the given force field models, the structural optimization was then repeated on an arbitrary structure of the same model system, in which the coordinates of every component atom were scrambled while the connection table was retained (Figure 4). All of the structural optimization computations were performed using the AMBER program. The Newton–Raphson method was applied to energy minimization. The convergence criterion was set to  $10^{-8}$  kcal/(mol·Å). The distance cutoff of nonbonded interactions was set to 999 Å. In each case, the root-mean-squared deviation (rmsd) of the resulting structure was calculated by using the structure preoptimized by QM as the reference. Only zinc and the four atoms in direct bonding with zinc were considered in rmsd calculations.

**Molecular Dynamics Simulations on Model Systems.** The five force field models were also applied to the molecular dynamics (MD) simulations of all four model systems in vacuum. All MD simulations were also conducted using the AMBER program. In each case, the preoptimized structure by QM was used as the starting structure for the following MD simulation. In order to release the internal strain energies of the entire system gradually, three rounds of restraint MD simulations were carried out first: (1) a 50 ps long simulation with restraints on nonhydrogen atoms (restraint harmonic force constant = 5.0 kcal/mol·Å<sup>2</sup>); (2) a 50 ps long simulation with restraints on nonhydrogen atoms (restraint harmonic force constant = 0.5 kcal/mol·Å<sup>2</sup>); and then (3) another 50 ps long simulation without any restraint. After these preparations, the final production run lasted for 10 ns, which was conducted in vacuum under a constant temperature of 300 K. The distance cutoff of nonbonded interaction was set to 999 Å. The periodic boundary condition was not enabled during simulation. The time interval was set to 1 fs during the entire simulation process, and the MD trajectory was also recorded every 1 ps for subsequent analyses.

Since the force field parameters derived in our study may be applied to the modeling of Zn-containing metalloproteins in their physiological environment, the five force field models (Table 1) were also applied to the MD simulations of all four model systems in water. In each case, the preoptimized structure by QM was soaked in a TIP3P water box<sup>44</sup> with a margin of 14 Å in each dimension. The entire system was

neutralized by adding an appropriate number of counterions, and a three-step minimization was used to release internal strain energies gradually. In each step, 5000 rounds of minimization was performed with the restraint harmonic force constant imposed on all nonhydrogen atoms set to 500.0, 10.0 kcal/mol·Å<sup>2</sup>, and zero, respectively. The entire system was then subjected to the same restraint MD simulation routine as the one performed in vacuum described in the previous paragraph. After all these preparative steps, a production simulation of 10 ns long was performed under constant temperature ( $T = 300$  K) and pressure ( $P = 1$  atm). Temperature of the entire system was regulated by Langevin thermostat<sup>45</sup> with the collision frequency  $\gamma = 2.0$  ps<sup>-1</sup>, and pressure of the system was controlled by Berendsen barostat.<sup>46</sup> The time interval was set to 1 fs. Periodic boundary condition was enabled during simulation. The distance cutoff for nonbonded interactions was 14 Å, and the particle mesh Ewald (PME) method<sup>47</sup> was used to compute long-range interactions. The MD trajectory was also recorded every 1 ps for subsequent analyses.

**2.4. Validation of the Derived Parameters on a Carbonic Anhydrase II Complex.** The force field parameters derived in our study were further validated on a complex structure formed by carbonic anhydrase II and 5-dimethylamino-naphthalene-1-sulfonamide (Figure 5). The complex structure was solved by Nair et al<sup>48</sup> through X-ray diffraction at a resolution of 2.10 Å (PDB entry: 1OKL). In this complex structure, the zinc ion inside the binding pocket is in coordination with three histidine residues (His90, His92, and His115) and with a sulfonamide moiety on the ligand molecule, a chemical configuration identical to *model system I*. Consequently, the force field parameters derived from *model system I* were applied in the following simulations.

Five separate MD simulations of this complex structure were performed to test force field models FF-1 to FF-5, respectively. To set up each simulation, the force field parameters for carbonic anhydrase II were taken from the AMBER FF03 parameter set, and those for the ligand molecule were taken from the AMBER GAFF parameter set. The van der Waals parameters for zinc were also set as:  $\sigma = 1.10$  Å and  $\epsilon = 0.0125$  kcal/mol. Note that, for all five force field models, the corresponding RESP charges or formal charges indicated in Figure 3 were applied only to the binding center, including the zinc ion, the imidazole moieties on His90/His92/His115, and the entire ligand

**Table 2.** Bond-Stretching and Angle-Bending Parameters Related to Zinc Derived from Four Zn-Containing Model Systems

bond type <sup>a</sup>	stretching force constant kcal/(mol·Å <sup>2</sup> )	equilibrium bond length (Å)
<i>Model System I</i>		
ZN-NB	56.0	2.07
ZN-n2	98.8	1.97
<i>Model System II</i>		
ZN-NB	49.5	2.09
ZN-s	92.8	2.26
<i>Model System III</i>		
ZN-NB	74.8	2.02
ZN-sh	33.2	2.48
<i>Model System IV</i>		
ZN-NB	62.9	2.33
ZN-o	98.8	1.97
ZN-s	62.9	2.34
bond angle <sup>a</sup>	bending force constant kcal/(mol·rad <sup>2</sup> )	equilibrium bond angle (°)
<i>Model System I</i>		
NB-ZN-NB	31.1	105.5
NB-ZN-n2	35.7	113.0
ZN-NB-CR	46.6	126.4
ZN-NB-CV	48.7	126.6
ZN-n2-hn	37.6	118.5
ZN-n2-s6	67.0	112.4
<i>Model System II</i>		
NB-ZN-NB	35.4	104.7
NB-ZN-s	27.6	113.8
ZN-NB-CR	49.4	126.2
ZN-NB-CV	49.4	127.0
ZN-s-c3	75.2	104.9
<i>Model System III</i>		
NB-ZN-NB	31.5	111.5
ZN-NB-CR	48.9	126.8
ZN-NB-CV	49.8	126.8
NB-ZN-sh	27.5	107.3
ZN-sh-hs	32.0	100.3
ZN-sh-c3	65.7	110.6
<i>Model System IV</i>		
CR-NB-ZN	43.2	123.21
CV-NB-ZN	43.9	129.78
NB-ZN-o	34.6	91.97
NB-ZN-s	32.7	97.52
o-ZN-s	26.3	113.89
s-ZN-s	21.6	129.12
c-o-ZN	60.7	119.82
c3-s-ZN	70.6	102.40

<sup>a</sup> The atom types used in this table are indicated in Figure 2.

molecule. The rest of the parts on the complex were assigned the template charges from the AMBER FF03 parameter set.

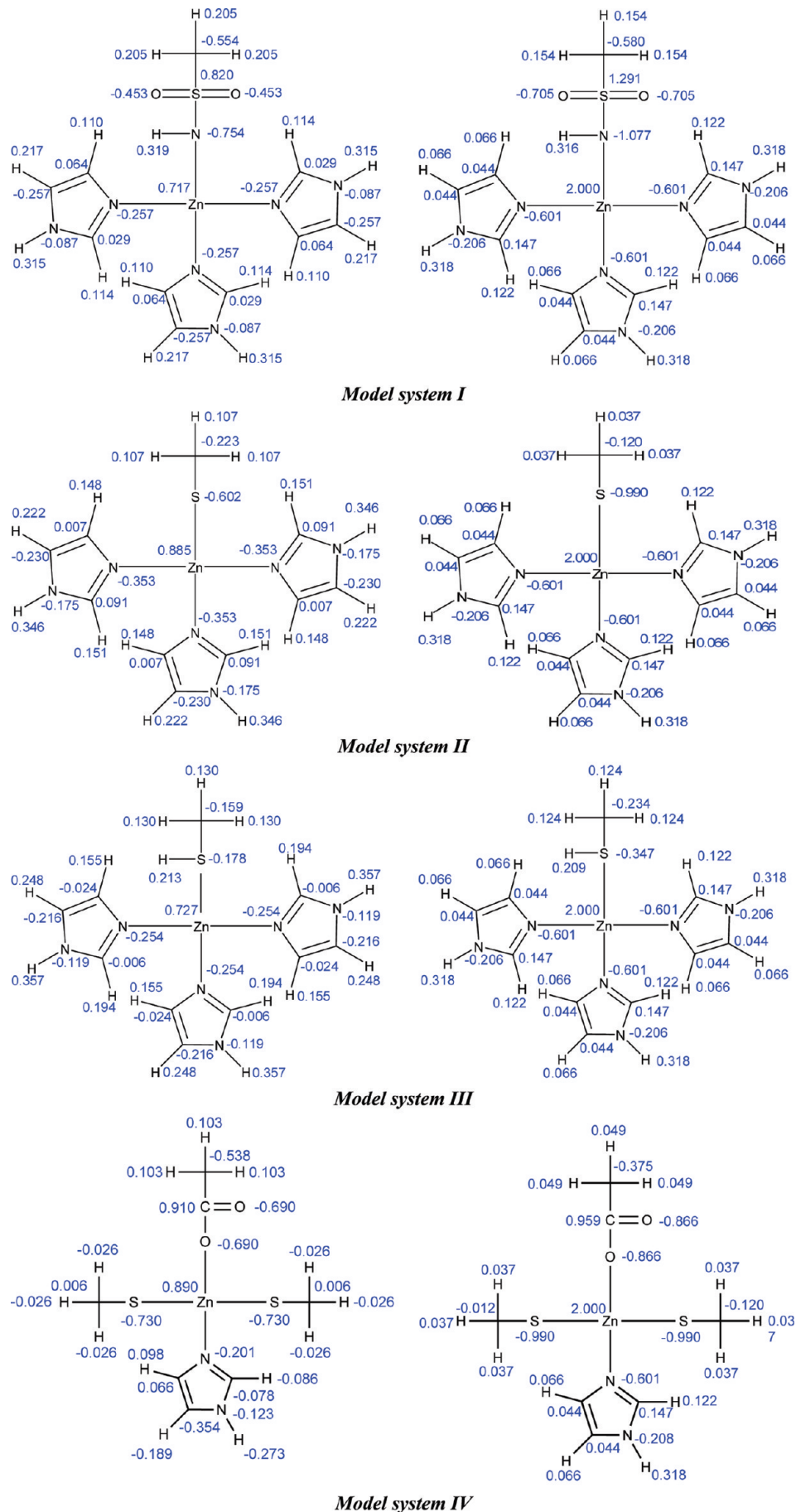
This complex was soaked in a TIP3P water box with a margin of 10 Å and was then neutralized by adding counterions. The subsequent stepwise minimization and restraint MD simulations were performed using the same procedure and settings as the MD simulations on Zn-containing model systems in water. After these preparative steps, a production simulation of 10 ns long was performed under a constant temperature ( $T = 300$  K) and a constant pressure ( $P = 1$  atm). Temperature of the system was also regulated by Langevin thermostat<sup>45</sup> with the collision

frequency  $\gamma = 2.0$  ps<sup>-1</sup>, and pressure of the system was controlled by Berendsen barostat.<sup>46</sup> The time interval for MD simulation was set to 2 fs. Periodic boundary condition was enabled during simulation. The distance cutoff for nonbonded interactions was 12 Å, and the PME method<sup>47</sup> was used to compute long-range interactions. In addition, the SHAKE algorithm<sup>49</sup> was applied to constrain all bonds involving hydrogen atoms. The MD trajectory was also recorded every 1 ps for subsequent analyses.

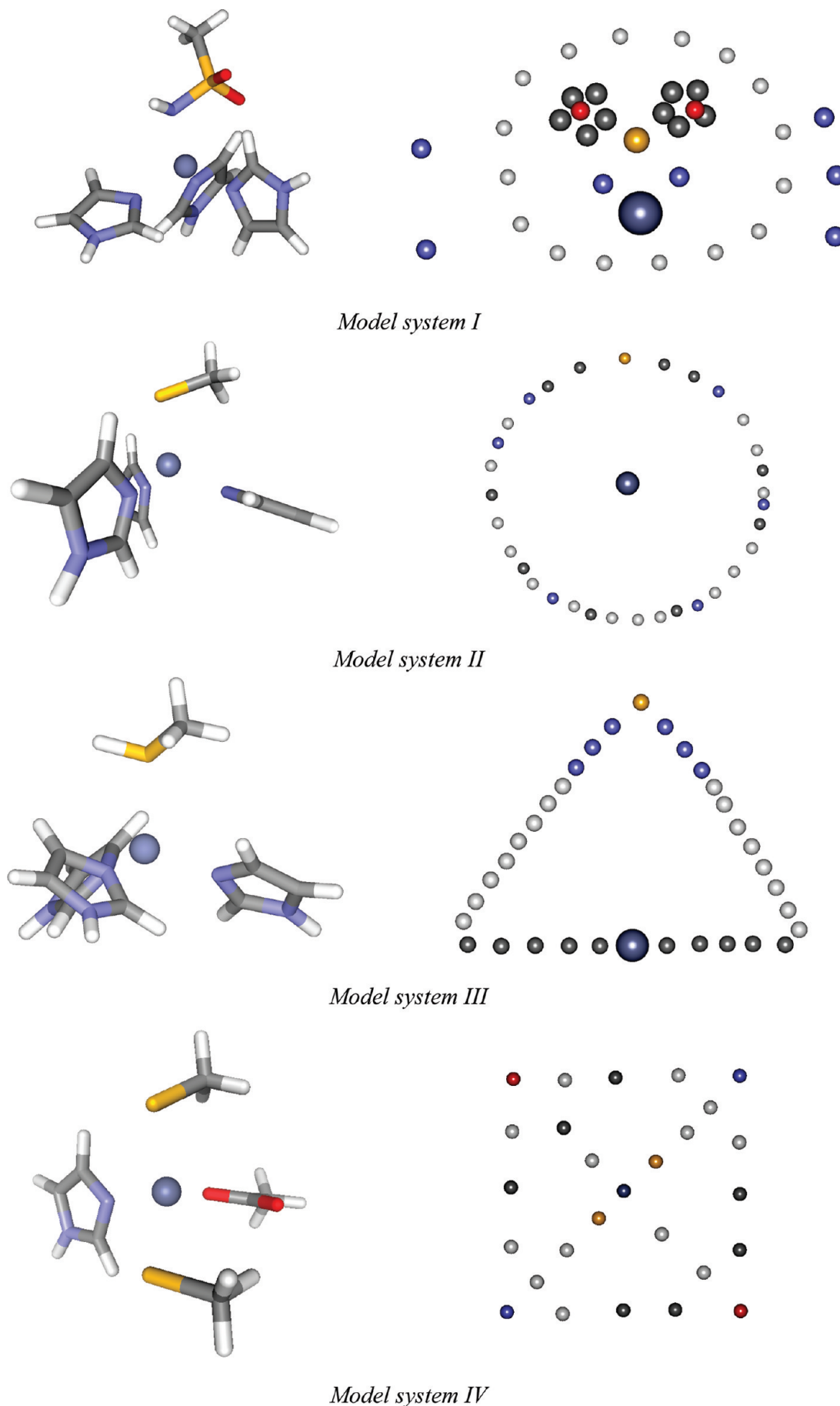
An additional MD simulation was performed on the same complex structure, in which the binding center was modeled by the QM/MM method<sup>50</sup> implemented in the AMBER program. The semiempirical PM3 method<sup>51</sup> was employed to treat the zinc ion, His90/His92/His115, and the entire ligand molecule. The SCF convergence was set to 10<sup>-8</sup> kcal/mol. The rest parts of the complex structure were still treated with the AMBER force field using the FF03 parameter set. The complex structure was also soaked in a TIP3P water box with a margin of 10 Å and was then neutralized by adding counterions. The entire system was subjected to the same stepwise minimizations and preparative restraint MD simulations as the other force field models. The final production run lasted for 4 ns under a constant temperature ( $T = 300$  K) and a constant pressure ( $P = 1$  atm). The time interval for MD simulation was set to 2 fs. The MD trajectory was recorded every 1 ps. All of the other major parameters/settings were the same as those used in the simulations by using other force field models.

**2.5. Further Optimization on Derived Parameters for Reproducing Vibrational Frequencies.** Producing the correct vibrational frequencies is also an important quality of a good force field model. For each model system illustrated in Figure 2, the vibrational frequencies computed at the B3LYP/6-311++G(2d,2p) level were compared with their counterparts given by AMBER with the FF-1 parameters in normal-mode analysis. These two sets of vibrational frequencies actually fit very well for all four model systems (Figure 6). Some obvious discrepancy was observed only at the high-frequency end (frequency >3000 cm<sup>-1</sup>). The normal modes given by QM computations on each model system were visually examined in the graphical user interface of the Gaussian 03 program to determine which were responsible for the observed discrepancy in vibrational frequencies. It turned out to be stretching of X-H bonds ( $X = C, N, O,$  or  $S$ ). Accordingly, the bond-stretching force constants of the X-H single bonds in FF-1 were further optimized to better reproduce the vibrational frequencies of *model systems I-IV*, and the outcomes were named as FF-5. All of the other parameters in FF-5 were the same as those in FF-1.

The optimization was carried out through a genetic algorithm (GA) procedure implemented in an in-house computer program. The optimization was carried out on a population of 500 chromosomes. Every chromosome was composed of a certain number of genes, each of which was encoded with the bond-stretching force constant of a particular X-H bond ( $X = C, N,$  or  $S$ ). The initial value of each gene was assigned the original value of the corresponding force constant in FF-1 plus a random perturbation within  $\pm 20\%$ . The fitness score of each chromosome was computed

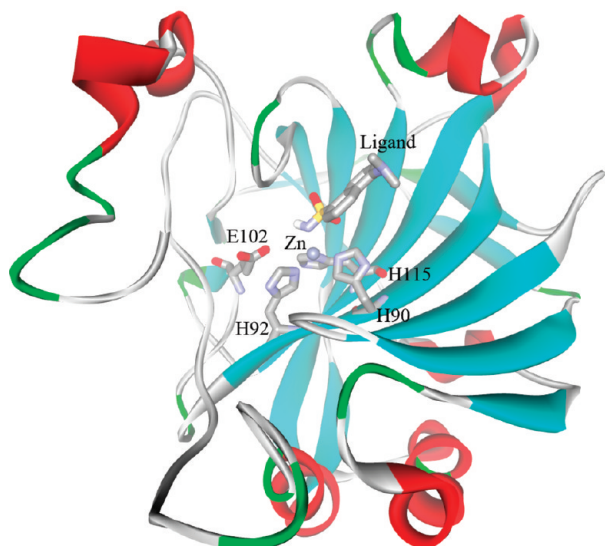


**Figure 3.** Atomic charges assigned on the four model systems when the RESP charge model was applied (left) or when a formal charge of +2e was applied to zinc (right).



**Figure 4.** Three-dimensional structures of *model systems I–IV* optimized at the B3LYP/6-311++G(2d, 2p) level (left), and the corresponding arbitrarily sketched structures used for validation (right).





**Figure 5.** Crystal structure of carbonic anhydrase II in complex with 5-dimethylamino-naphthalene-1-sulfonamide (PDB entry: 1OKL), which was used for the validation of force field parameters in this study.

as the rmsd between the vibration frequencies given by QM computations and those given by the normal-mode analysis in AMBER applying the force constants encoded in the given chromosome. Thus, the smaller was its fitness score, the better was the given chromosome. The entire population undertook optimization in the steady-state mode<sup>52</sup> for a total of 50 000 GA rounds. At each round, two types of genetic operations, including single-point mutation and crossover, may occur at a probability of 30% and 70%, respectively. The single-point mutation occurred randomly at a particular gene on a given parent chromosome in which that gene was altered randomly within  $\pm 20\%$  of its original value. The single-point crossover occurred between two parent chromosomes in which the two chromosomes exchanged their genes starting from a random point. For both the mutation and the crossover operations, the parent chromosomes were selected using the roulette-wheel method,<sup>52</sup> so that better chromosomes had a greater chance to produce offsprings. At each GA round, the newly generated chromosomes (one or two) were compared with the worst chromosome in the entire population. If the new one had better fitness scores, then they would replace the worst chromosome in the population; otherwise the population would remain as the same to enter next GA round. The average fitness score of the entire population was monitored along the entire GA process. For all four model systems, the average fitness score of the entire population actually reached convergence well before 50 000 GA steps. After the GA process was completed, the force constants encoded in the best chromosome were retrieved as the final parameters in FF-5.

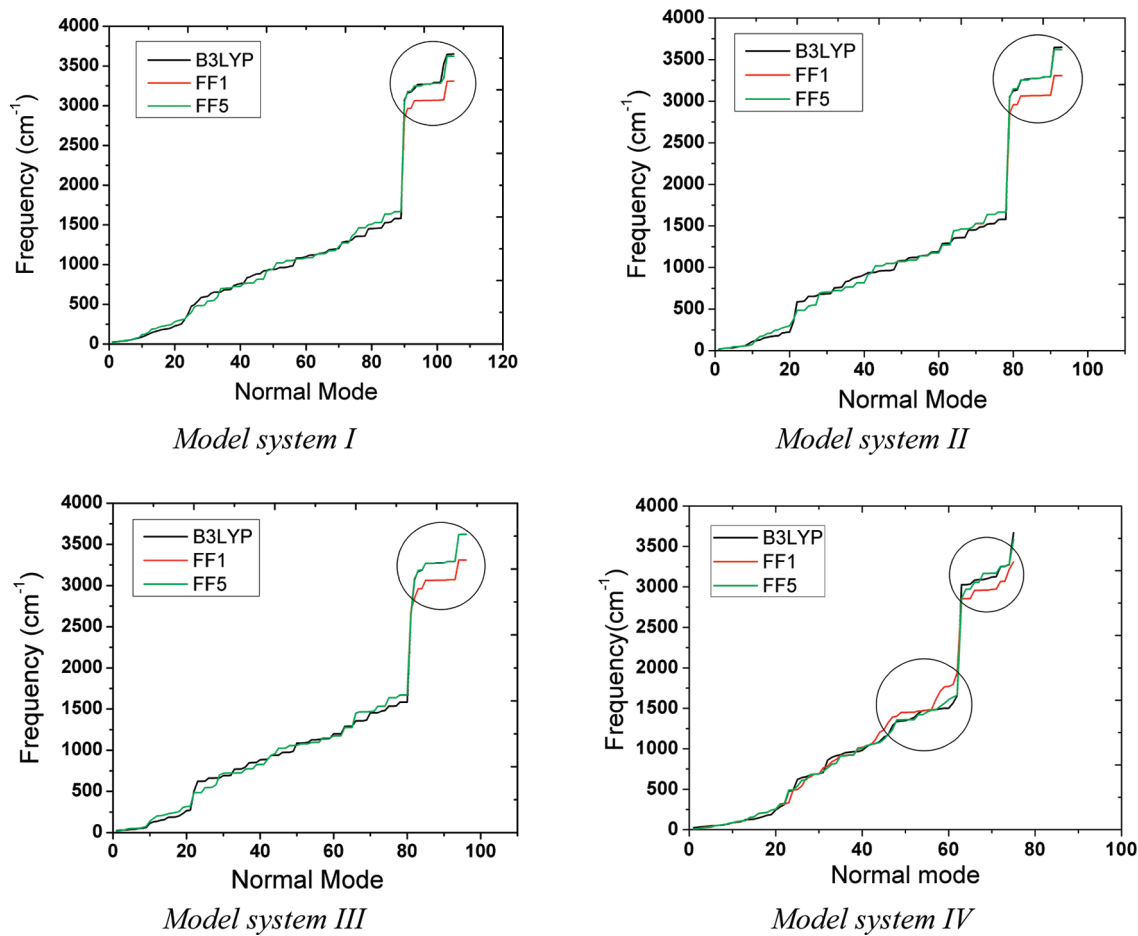
To make comparison with other force field models, FF-5 was also applied to the structural optimizations and MD simulations on *model systems* I–IV as well as the MD simulations on the carbonic anhydrase II complex, using the same settings described earlier in this manuscript.

### 3. Results and Discussion

**3.1. Validation of Force Field Models on Four Simple Model Systems.** As described in the Methods Section, all five force field models (FF-1 to FF-5) were applied to the structural optimizations of *model systems* I–IV (Figure 2) in both vacuum and water. The rmsd values between the structures optimized by FF-1 to FF-5 and the corresponding structures optimized at the B3LYP/6-311++G(2d, 2p) level are summarized in Table 1. As one example, the three-dimensional structures of *model system* I optimized by these force field models in vacuum are illustrated in Figure 7.

One can see that FF-1 (bonded model + RESP charges) performed very well in reproducing the correct three-dimensional structures of all four model systems, no matter if the structural optimization was started from a preoptimized structure or a ridiculous structure. The rmsd values of the structures optimized by FF-1 are generally below 0.25 Å on all four model systems. In contrast, the performance of FF-2 (bonded model + formal charges) was generally inferior to that of FF-1. For example, FF-2 produced a relatively large rmsd value in handling the structure of *model system* I in both vacuum and water. This may be attributed to the strong secondary electrostatic interactions between the zinc ion (atomic charge = +2.000e) and the two oxygen atoms (atomic charge = -0.705e) on the sulfonamide moiety on the ligand molecule. Indeed, one can see in the structure produced by FF-2 (Figure 7) that the zinc ion tends to get closer to these two oxygen atoms and results in an obvious distortion of the tetrahedral geometry of the coordination center. Another disadvantage of FF-2 is that, unlike FF-1, its performance is somewhat unpredictable; its performance was not so good on *model system* I; whereas it was acceptable on *model system* II. As for *model system* III, in which the ligand molecule is in the neutral form, it produced reasonable structures in vacuum but not in water.

As for the two nonbonded models, FF-4 (nonbonded model + formal charges) demonstrated a relatively robust performance across all four model systems. The rmsd values of the structures produced by FF-4 are generally between 0.2–0.4 Å no matter if the optimization was performed in vacuum or water. This level of accuracy, however, is still inferior to the one produced by FF-1. The performance of FF-3 (nonbonded model + RESP charges) was not consistent across four model systems; it produced reasonable structures of *model system* II, whereas it had some obvious problems in reproducing the structures of I, III, and IV in vacuum. In particular, for *model system* III, in which the ligand molecule is in the neutral form, the electrostatic interaction between the zinc ion (charge = +0.727e) and the sulfur atom (charge = -0.178e) on the thiol group is not strong enough for maintaining these two atoms in a bonding range. The structure of this model system was basically disrupted upon optimization with FF-3, resulting in a large rmsd value of over 16 Å. These results suggest that nonbonded models, which rely only on electrostatic and van der Waals interactions for modeling the coordination configuration of zinc, are less capable than bonded models. If a nonbonded model



**Figure 6.** Vibrational frequencies of four model systems computed at the B3LYP/6-311++G(2d, 2p) level (black), force field model FF-1 (red), and the specially optimized force field model FF-5 (green). In each figure, the normal modes are sorted by their vibrational frequencies computed by B3LYP as the x-axis.

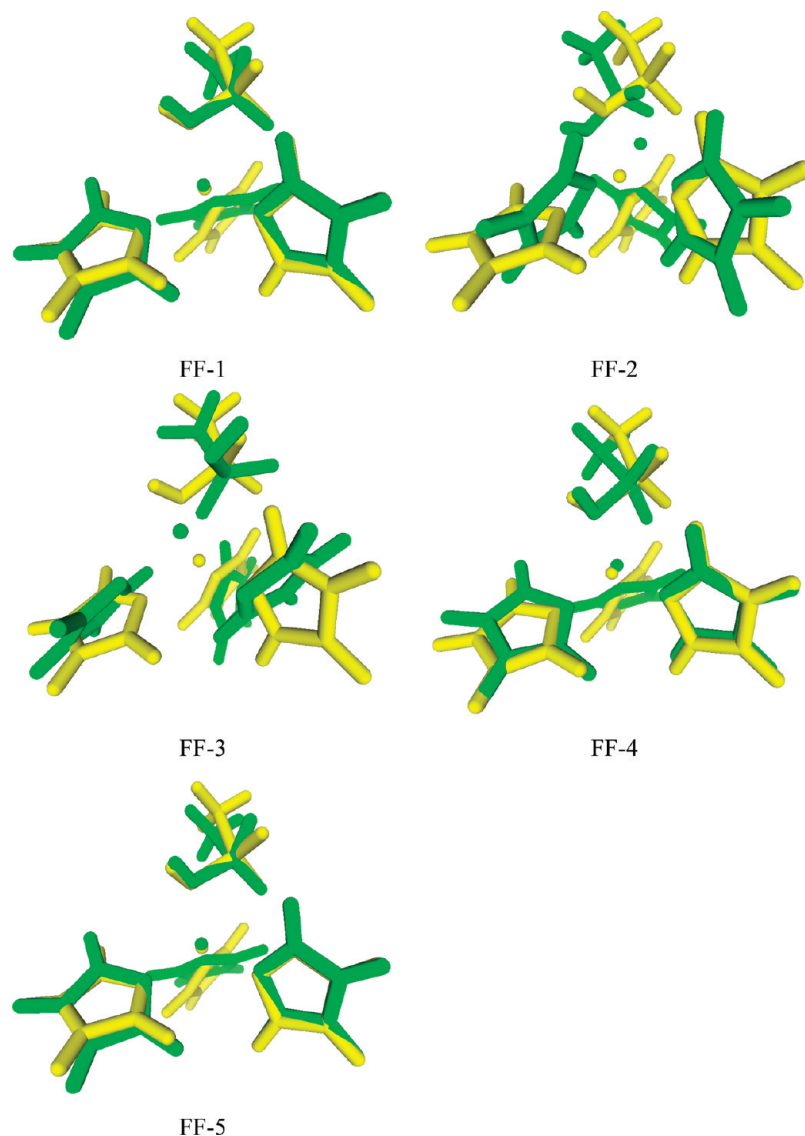
is chosen for this purpose anyway, application of RESP charges will not improve its performance.

The five force field models were also tested in extensive MD simulations on the four Zn-containing model systems. When the MD simulations were performed in vacuum, all force field models except FF-3 were able to basically maintain the stable structures of all four model systems (data not shown). In the case of FF-3, the structures under simulation went complete wrong rapidly (<20 ps). The rmsd curves monitored along the entire MD trajectory produced by FF-1 to FF-5 in explicit water are illustrated in Figure 8. One can see clearly that both nonbonded models (FF-3 and FF-4) were not able to maintain the stable structures of all four model systems. Note that the oxygen atom in a TIP3P water molecule carries a substantial amount of partial charge ( $-0.834e$ ). The water molecule thus acts as a strong competitor for bonding with the zinc ion, which could be challenging for nonbonded models for maintaining the desired coordination configuration of zinc. Bonded models have obvious advantages in this aspect. As revealed in Figure 8, the desired tetrahedral coordination configuration of zinc was highly stable during the entire simulation by FF-1. As for FF-2, the four model systems underwent some noticeable structural fluctuations from time to time. This may also be attributed to the exaggerated secondary electrostatic interac-

tions between the zinc ion and the water molecules due to the formal charge assigned on the zinc ion ( $+2e$ ).

The results observed in the structural optimizations and the MD simulations of the four model systems are basically consistent. They both indicate that the performance of bonded models (FF-1 and FF-2) is superior to that of nonbonded models (FF-3 and FF-4) in reproducing the desired coordination configuration of zinc. Note that the major difference between FF-1 and FF-2 is whether to employ the RESP or the formal charge model in the computation of electrostatic interactions. The formal charge model assumes that a formal charge of  $+2e$  is localized on the zinc ion, and the rest of the parts are not affected by it. In contrast, the charge carried by the zinc ion is dispersed on the entire system according to the RESP model, which better mimics the charge transfer effect between zinc and the surrounding chemical moieties. As indicated by our results, FF-1 is generally more accurate and more robust than FF-2. We thus conclude that FF-1 (bonded model + RESP charges) is the better choice for modeling Zn-containing molecular systems.

**3.2. MD Simulation of a Carbonic Anhydrase II Complex in Explicit Solvent.** Our force field parameters are derived from some simple model systems. The really meaningful application of these parameters will be the modeling of the complexes formed by Zn-containing met-



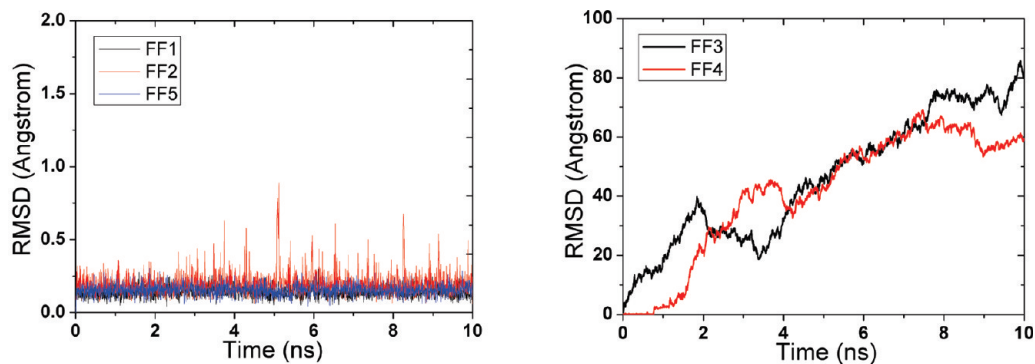
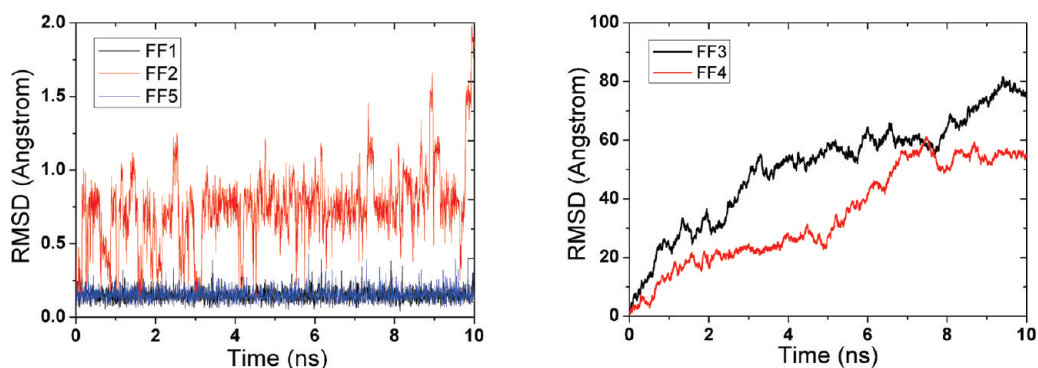
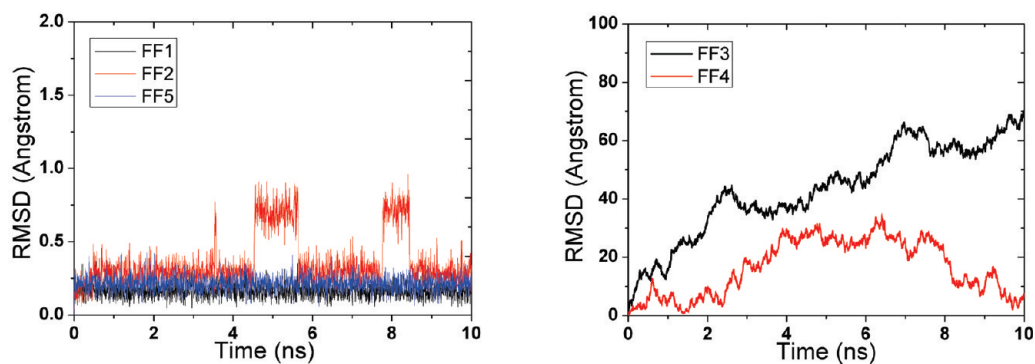
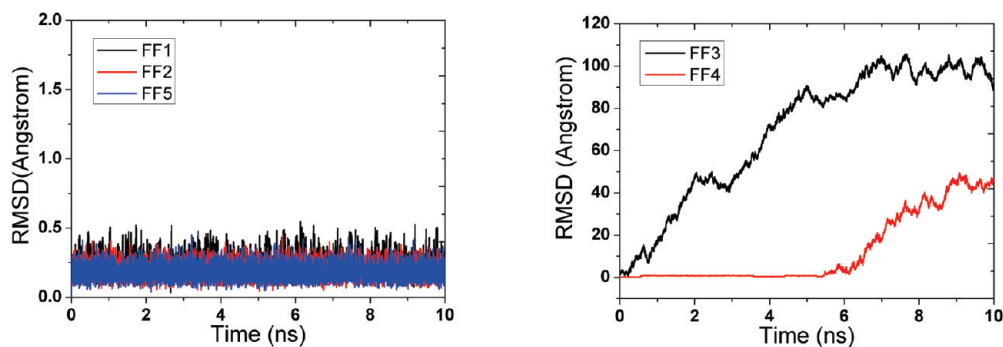
**Figure 7.** Three-dimensional structures of *model system I* optimized by FF-1 to FF-5 in vacuum based on a ridiculous initial structure. These structures are superimposed with the one optimized at the B3LYP/6-311++G(2d, 2p) level (the structure in yellow).

alloproteins. Thus, we have chosen a protein–ligand complex formed by carbonic anhydrase II, a typical Zn-containing enzyme with pharmaceutical implications, to further validate all five force field models. In this complex structure, the zinc ion inside the binding pocket is in coordination with three histidine residues on the protein (His90, His92, and His115) and a sulfonamide group on the ligand molecule, a chemical configuration identical to *model system I*. In addition, an interesting feature of this complex is that Glu102 is near the coordination center (Figure 5), which is not in direct bonding with the zinc ion (Zn–O distance = 3.9 Å). This feature makes the complex more challenging for simulation since the negatively charged side chain of Glu102 could disrupt the desired coordination configuration of the zinc ion inside the binding pocket.

The five force field models were applied to the MD simulations of this complex structure in explicit water. The force field parameters derived from *model system I* were applied to the simulations by the bonded models, i.e., FF-1, FF-2, and FF-5. Rmsd values of the coordination center (the

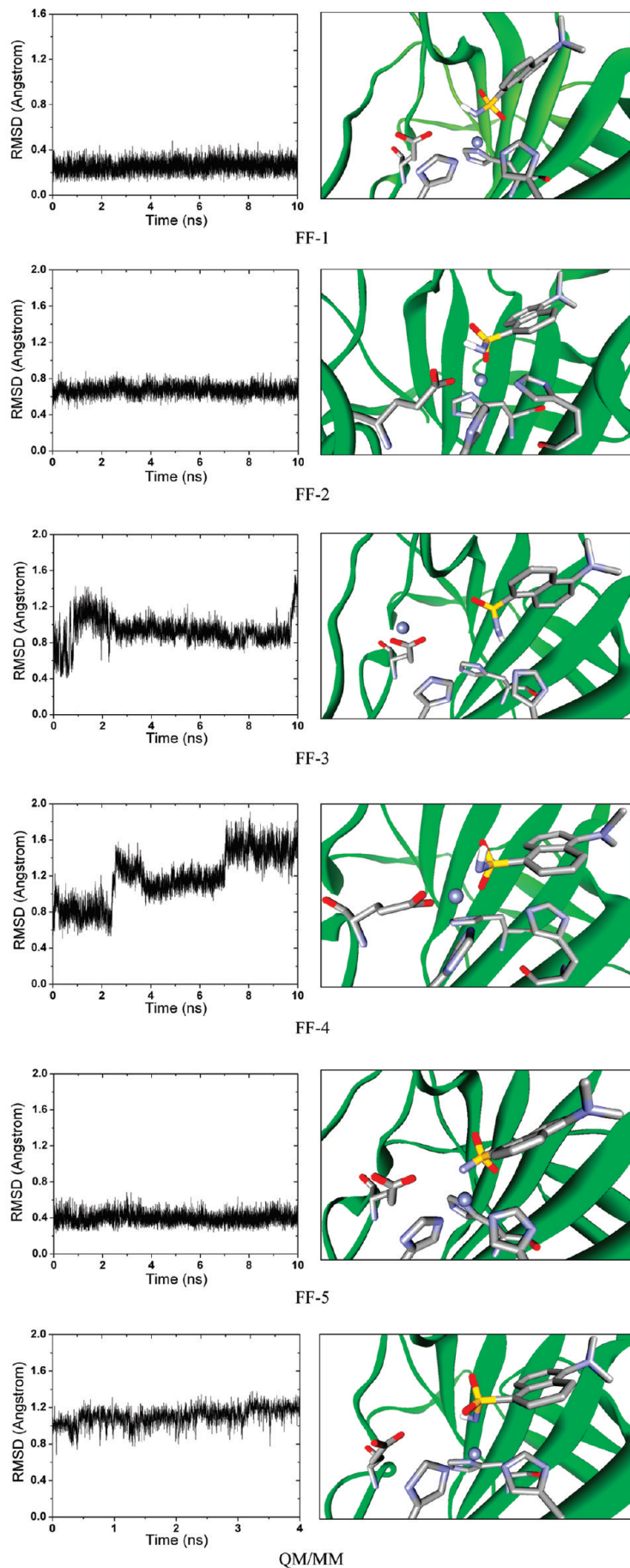
zinc ion plus four atoms in direct coordination with it) were monitored along each MD trajectory (Figure 9). One can see that these three bonded models were able to maintain the tetrahedral geometry of the coordination center very well during the entire simulation. In particular, the average rmsd values produced by FF-1 and FF-5 are as small as  $\sim 0.4$  Å. The average rmsd values produced by FF-2 ( $\sim 0.7$  Å) are slightly larger than those of FF-1 and FF-5. In the last snapshot on the MD trajectory of FF-2, the carboxyl group on the side chain of Glu102 tends to get closer to the zinc ion as compared to the structures produced by FF-1 and FF-5. In fact, one of the oxygen atoms on that carboxyl group is already in a bonding range with the zinc ion (Zn–O distance = 1.80 Å) in this structure. This observation further proves that the significant formal charge assigned on the zinc ion is not completely reasonable.

As for the two nonbonded models (FF-3 and FF-4), relatively large rmsd values are observed on their MD trajectories (Figure 9). One can see clearly in the structures produced by FF-3 and FF-4 that the zinc ion has moved out

*Model system I**Model system II**Model system III**Model system IV*

**Figure 8.** Rmsd values monitored along the MD trajectories produced by five force field models in explicit water. In each case, the rmsd values are computed by considering only the coordination center, while the structure optimized at the B3LYP/6-311++G(2d, 2p) level is used as the reference.





**Figure 9.** Rmsd values of the coordination center on a carbonic anhydrase II complex structure monitored along each MD trajectory (left), and the last snapshot retrieved from each MD trajectory (right) produced by six different methods.

the coordination center and is now close to the carboxyl group on the side chain of Glu102. Consequently, the tetrahedral geometry of the coordination center is disrupted considerably. It seems that the complex structure under simulation has not reached equilibrium at 10 ns, and it would undergo even more significant changes if the simulation was extended. The potential problem of nonbonded models in modeling Zn-containing protein–ligand complexes is clearly demonstrated in this test.

Recently, Lim et al developed a nonbonded force field model with a potential energy function, including terms for charge transfer and polarization effects.<sup>13</sup> They conducted molecular dynamics simulations of Zn<sup>2+</sup> bound to Cys and His residues in proteins using both conventional force field and their modified model. In their study, simulations with the conventional force field yielded a nontetrahedral Cys<sub>2</sub>His<sub>2</sub> Zn-binding configuration and significantly overestimated the experimental Zn–S bond length. In contrast, simulations with their new potential energy function better reproduced the experimentally observed tetrahedral Cys<sub>2</sub>His<sub>2</sub> and Cys<sub>4</sub> Zn-binding configurations. Lim's study is another good example for demonstrating the limitations of conventional force field models in the simulation of Zn-containing molecular systems. Some appropriate considerations on charge transfer and polarization effect are thus desired if nonbonded force field models are to be used to study such systems.

Nevertheless, in Lim's study the charge transfer was restricted between the zinc ion and the atoms in direct bonding with it, and the amount of transferred charge needed to be recomputed at each time step during simulation. More importantly, adding a polarization term into an existing force field model certainly requires careful reparameterizations and validations. It can be fairly complicated especially when diverse chemical structures have to be considered. In contrast, our approach only needs to define a number of new bond types and to supply the necessary parameters, which is technically more practical. In this study, we have demonstrated the application of this approach to a variety of ligand molecules bound with zinc. It was also unclear in Lim's study if their new potential energy function could reproduce vibrational frequencies, a more challenging goal for force field models. As described later in this article, the force field parameters derived in our study are capable for this purpose.

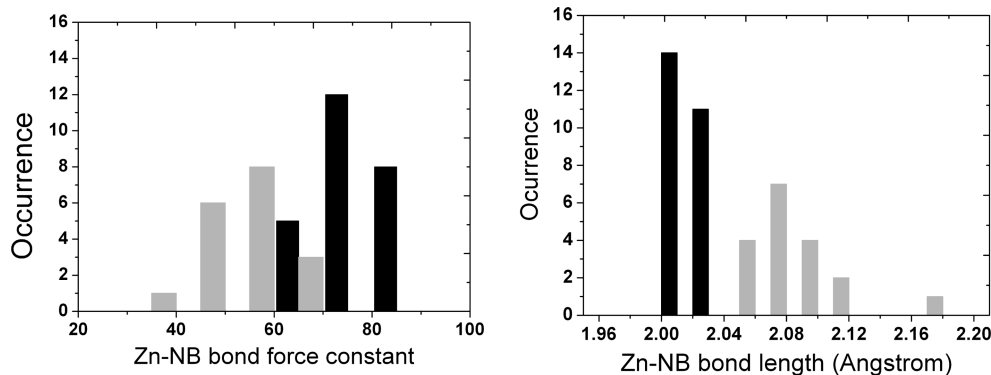
Besides force field models, the QM/MM method implemented in the AMBER program was also applied in our study to modeling of the same carbonic anhydrase II complex. We found that the MD simulation by employing QM/MM was about five times slower than the MD simulation by employing the AMBER force field. Thus, the MD simulation with QM/MM was performed only for 4 ns due to the significant computation cost. The tetrahedral geometry of the zinc coordination center is basically maintained during this simulation (Figure 9). It, however, undergoes a notable distortion from the original crystal structure (rmsd  $\sim$  1.3 Å). The same trend is also indicated by the rmsd values computed considering the entire ligand molecule (see the Supporting Information), where the structure produced by the QM/MM simulation exhibits a greater change compared to those produced by FF-1 and FF-5.

It is interesting to find that the force field parameters derived in our study (FF-1/FF-5) outperformed the QM/MM method in modeling this carbonic anhydrase complex. It prompts that, although the concept of QM/MM is appealing, this method is not automatically more accurate than a conventional MM approach. The less encouraging performance of QM/MM observed in our study may be attributed to some technical reasons. For example, the QM computations in QM/MM actually employed the semiempirical PM3 method, which may not produce satisfactory results in this particular case. There are also some other parameters which may affect the final outcomes of the QM/MM method. An optimal set of these parameters may lead to better outcomes. Nevertheless, exploring the QM/MM method implemented in the AMBER program is certainly beyond the scope of this study.

**3.3. Reproduction of the Vibrational Frequencies of Zinc-Containing Model Systems.** Besides three-dimensional structures, producing the correct vibrational frequencies is also a desired quality of a force field. The vibrational frequencies of four model systems computed at the B3LYP/6-311++G(2d, 2p) level and the force field models FF-1 and FF-5 are illustrated in Figure 6. One can see that, although most vibrational frequencies produced by FF-1 match well with their counterparts produced by QM computations, some notable discrepancies at the high-frequency end ( $>3000\text{ cm}^{-1}$ ) still exist. As described in the Methods Section, we found that stretching motions of the X–H bonds (X = C, N, or S) were largely responsible for this. Consequently, the bond-stretching force constants of such bonds in FF-1 were further optimized through a genetic algorithm approach to better reproduce vibrational frequencies. The resulting parameters, named as FF-5, are tabulated in the Supporting Information together with the original ones in FF-1 for all four model systems.

FF-5 was also applied to optimize the structures of all four model systems. In each case, the structure preoptimized at the B3LYP/6-311++G(2d,2p) level was used as the starting structure, and the final optimized structure was subjected to normal model analysis by using the AMBER program. The vibrational frequencies computed thereby are compared with those produced by QM computations in Figure 6. One can see that FF-5 indeed reproduces the vibrational frequencies at the high-frequency end very well on all four model systems. This observation indicates that our method for optimizing FF-1 is effective; the desired goal can be achieved simply by adjusting the stretching force constants of the X–H bonds within a reasonable range ( $\pm 20\%$ ) of their original values.

Thus, FF5 is more accurate than FF1 in terms of producing correct vibrational frequencies. The only difference between FF1 and FF5 lies in some bond-stretching force constants of X–H bonds. Thus, it is not surprising that, as demonstrated in all of our tests, FF1 and FF5 are equally capable of reproducing the correct three-dimensional structures of simple Zn-containing model systems as well as a Zn-containing protein–ligand complex. In fact, accurate modeling of high-frequency vibrations is rarely a matter of concern for modeling biological macromolecules. For example, the



**Figure 10.** Distribution of the bonding stretching parameters for the Zn–NB bond derived from 18 typical Zn-containing model systems. Columns in black and gray represent the parameters for model systems in which the ligands are neutral and deprotonated, respectively.

popular SHAKE approximation in the AMBER program fixes the lengths of all X–H bonds during MD simulation in order to reduce computational costs. Thus, one can safely apply FF-1 instead to Zn-containing metalloproteins, saving an extra amount of efforts on parameter optimization.

**3.4. On the Derived Force Field Parameters.** A notable feature of our study is that we constructed a variety of Zn-containing model systems and then applied Seminario's method to derive force field parameters for each of them. As described in the Methods Section, an extensive survey was performed on the entire PDB to identify common chemical moieties in bonding with zinc, including acid, ketone, sulfone, alcohol, thiol, amide, and amine groups. The model systems summarized in Figure 1 are designed to mimic the binding of these chemical groups with Zn-containing proteins. Note that, for alcohol and thiol groups, both of their neutral and deprotonated forms are considered (M5/M6, M7/M8, and M9/M10) in order to explore the possible difference in their bonding with zinc. In addition, six more model systems (M-13 to M-18) are used to mimic some mixed zinc coordination centers observed in Zn-containing proteins. The complete list of the bond-stretching and angle-bending parameters derived in our study for all 18 Zn-containing model systems are summarized in the Supporting Information. They are readily applicable to the molecular modeling studies of Zn-containing proteins, either in bound or unbound states, with the AMBER program.

We have observed in our results that the force field parameters for the same type of bond or angle derived from different model systems are not identical. Instead, they scatter in a certain range. Since most model systems considered in our study contain three imidazole moieties in binding with zinc, we use the Zn–NB bond (NB is a nitrogen atom on the imidazole ring, see Figure 2) here as an example to illustrate this issue. Distributions of the parameters for the Zn–NB bond derived on all model systems are given in Figure 10. One can see that the equilibrium bond lengths of the Zn–NB bond are between 2.00–2.18 Å, while the corresponding stretching force constants are between 30–90 kcal/(mol·Å<sup>2</sup>). In fact, different sets of these parameters are also reported in literature. For example, the corresponding data reported by Lu et al.<sup>19</sup> were 2.27 Å and 26.0 kcal/(mol·Å<sup>2</sup>), the data reported by Tuccinardi et al.<sup>20</sup> were 2.08 Å and 99.0 kcal/(mol·Å<sup>2</sup>), and the data derived from the

Raman spectrum of [Zn(NH<sub>3</sub>)<sub>4</sub>]<sub>2</sub> crystals were 2.10 Å and 40.0 kcal/(mol·Å<sup>2</sup>).<sup>19,53</sup>

In particular, one can see the parameters of the Zn–NB bond shown in Figure 10 can be divided into two distinct groups. The Zn–NB bond tends to be shorter (2.00–2.04 Å) and its stretching force constant is larger (60–90 kcal/mol·Å<sup>2</sup>) when the ligand molecule binding with zinc is in its neutral form. In contrast, the Zn–NB bond tends to be longer (2.04–2.18 Å) and its stretching force constant is smaller (30–70 kcal/mol·Å<sup>2</sup>) when the ligand molecule is in its deprotonated form. This can be easily understood: when the ligand is deprotonated, i.e., negatively charged, the bonding between the zinc ion and the ligand is stronger and is associated with a shorter bond length. Consequently, the bonds between zinc and imidazole rings are weakened and become slightly longer. This actually can be proven by comparing the relevant bond-stretching parameters derived from *model systems* II and III (Table 2). The Zn–S bond between zinc and the –SH group is longer than the counterpart between zinc and –S<sup>–</sup> by 0.22 Å, and it is much weaker than the latter. Results derived from other model systems in a neutral/deprotonated pair also reveal the same trend (see the Supporting Information).

Our results, together with the previous results reported by other researchers, suggest that zinc-related force field parameters are not always transferable across different systems. As revealed in some previous studies,<sup>19,20</sup> parameters originally derived from other model systems had to be optimized on the specific system under study through trial-and-error efforts; otherwise it could be risky. For this reason, we have considered a variety of Zn-containing model systems and produced a more comprehensive set of parameters. When one wants to model a specific Zn-containing system, he/she may choose and apply the appropriate force field parameters derived from a corresponding model system, which will in turn reduce the efforts on parameter optimization and produce more accurate results. This is a noteworthy advantage of our study as compared to previous studies.

Another noteworthy advantage of our study, which is perhaps more important, is that we have demonstrated a complete procedure for the deduction and validation of force field parameters. We have chosen Seminario's method for deriving force field parameters from the outcomes of QM computation, which has obvious technical advantages. First,



it is based on a Hessian matrix in the Cartesian coordinates. Thus, it avoids the troubles in setting internal coordinates as seen in some previous studies.<sup>25</sup> Second, the desired force constants of multiple bonds and bond angles can be derived simultaneously in one job rather than through an iterative trial-and-error approach. Very little human inference is needed during the whole process. In fact, once the Hessian matrix of a given model system is available, the rest of the steps can be largely automated by computer programs. Due to this advantage, we were able to process a variety of model systems and obtain the desired parameters. Application of Seminario's method is not limited to the 18 selected Zn-containing systems considered in this study. The same approach certainly can be employed to handle other Zn-containing molecular systems whenever necessary. It, in principle, can be applied to other types of metal ions or uncommon chemical moieties lacking appropriate force field parameters as well.

#### 4. Conclusions

In this study, we have derived bond-stretching and angle-bending parameters applicable to zinc-containing systems through a systematic approach. A total of 18 Zn-containing model systems were considered, and Seminario's method was applied to analyze the Hessian matrix of each model system computed at the B3LYP/6-311++G(2d,2p) level to derive the desired force constants. Then, the derived parameters were validated extensively in structural optimization and molecular dynamics simulations of four model systems as well as one protein–ligand complex formed by carbonic anhydrase II. With the application of these parameters, the bonded model in combination with the RESP charges (FF-1) was founded to be most robust in reproducing the three-dimensional structures of these Zn-containing systems; whereas the performance of nonbonded models was generally inferior. The performance of FF-1 was even better than the quantum mechanics/molecular mechanics (QM/MM) method implemented in the AMBER program in the MD simulations of a carbonic anhydrase II complex in explicit water. After some necessary optimizations on the force constants of X–H bonds, i.e., FF-5, it was also able to reproduce the vibrational frequencies of each model system provided by QM computations. Thus, the force field parameters derived in our study seem to be very reliable. Our approach, which is based on Seminario's method, has certain technical advantages. It can derive the parameters for all relevant chemical bonds in one batch rather than through an iterative procedure. Application of this approach is certainly not limited to the Zn-containing systems considered in this study. It is, in principle, applicable to molecular systems containing other types of metal ions or uncommon chemical moieties in which appropriate force field parameters are currently not available.

**Acknowledgment.** The authors are grateful to the financial supports from the Chinese National Natural Science Foundation (grant no. 20772149 and 90813006), the Chinese Ministry of Science and Technology (grant no. 2006AA02Z337 and 2009ZX09501-002), and the Science and Technology Commission of Shanghai Municipality (grant no. 074319113).

The authors also thank Dr. Eddy Bautista at Texas A&M University in College Station, Texas and Prof. Ulf Ryde at Lund University in Sweden for their helpful instructions on Seminario's method.

**Supporting Information Available:** A complete list of the force field parameters derived in this study. This material is available free of charge via the Internet at <http://pubs.acs.org/>.

#### References

- (1) Sarkar, B. Metal protein interactions. *Prog. Food Nutr. Sci.* **1987**, *11*, 363–400.
- (2) Dudev, T.; Lin, Y. L.; Dudev, M.; Lim, G. First-Second Shell Interactions in Metal Binding Sites in Proteins: A PDB Survey and DFT/CDM Calculations. *J. Am. Chem. Soc.* **2003**, *125*, 3168–3180.
- (3) Auld, D. S. Zinc coordination sphere in biochemical zinc sites. *BioMetals* **2001**, *14*, 271–313.
- (4) Lipscomb, W. N.; Strater, N. Recent Advances in Zinc Enzymology. *Chem. Rev.* **1996**, *96*, 2375–2433.
- (5) Brinckerhoff, C. E.; Matrisian, L. M. Matrix metalloproteinases: a tail of a frog that became a prince. *Nat. Rev. Mol. Cell Biol.* **2002**, *3*, 207–214.
- (6) Tallant, C.; Marrero, A.; Gomis-Rüth, F. X. Matrix metalloproteinases: Fold and function of their catalytic domains. *Biochim. Biophys. Acta* **2010**, *1803*, 20–28.
- (7) Ryde, U. Combined quantum and molecular mechanics calculations on metalloproteins. *Curr. Opin. Chem. Biol.* **2003**, *7*, 136–142.
- (8) Ryde, U. The coordination of the catalytic zinc ion in alcohol dehydrogenase studied by combined quantum-chemical and molecular mechanics calculations. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 153–164.
- (9) Brancato, G.; Rega, N.; Barone, V. Microsolvation of the Zn(II) ion in aqueous solution: A hybrid QM/MM MD approach using non-periodic boundary conditions. *Chem. Phys. Lett.* **2008**, *451*, 53–57.
- (10) Hartsough, D. S.; Merz, K. M. Dynamic force field models: Molecular dynamics simulations of human carbonic anhydrase II using a quantum mechanical/molecular mechanical coupled potential. *J. Phys. Chem.* **1995**, *99*, 11266–11275.
- (11) Stote, R. H.; Karplus, M. Zinc Binding in Proteins and Solution: A Simple but Accurate Nonbonded Representation. *Proteins: Struct., Funct., Genet.* **1995**, *23*, 12–31.
- (12) Donini, O. A.; Kollman, P. A. Calculation and prediction of binding free energies for the matrix metalloproteinases. *J. Med. Chem.* **2000**, *43*, 4180–4188.
- (13) Sakharov, D. V.; Lim, C. Zn Protein Simulations Including Charge Transfer and Local Polarization Effects. *J. Am. Chem. Soc.* **2005**, *127*, 4921–4929.
- (14) Pang, Y. P. Novel zinc protein molecular dynamics simulations: Steps toward antiangiogenesis for cancer treatment. *J. Mol. Model.* **1999**, *5*, 196–202.
- (15) Pang, Y. P.; Xu, K.; El Yazal, J.; Prendergast, F. G. Successful molecular dynamics simulation of the zinc-bound farnesyltransferase using the cationic dummy atom approach. *Protein Sci.* **2000**, *9*, 1857–1865.
- (16) Vendani, A.; Huhta, D. W. A New Force Field for Modeling Metalloproteins. *J. Am. Chem. Soc.* **1990**, *112*, 4759–4767.



- (17) Merz, K. M. Insights into the function of the zinc hydroxide-Thr199-Glu106 hydrogen bonding network in carbonic anhydrases. *J. Mol. Biol.* **1990**, *214*, 799–802.
- (18) Suarez, D.; Merz, K. M. Molecular Dynamics Simulations of the Mononuclear Zinc- $\beta$ -lactamase from *Bacillus Cereus*. *J. Am. Chem. Soc.* **2001**, *123*, 3759–3770.
- (19) Lu, Q.; Tan, Y. H.; Luo, R. Molecular Dynamics Simulations of p53 DNA-Binding Domain. *J. Phys. Chem. B* **2007**, *111*, 11538–11545.
- (20) Tuccinardi, T.; Martinelli, A.; Nuti, E.; Carelli, P.; Balzano, F.; Uccello-Barretta, G.; Murphy, G.; Rossello, A. Amber force field implementation, molecular modelling study, synthesis and MMP-1/MMP-2 inhibition profile of (R)- and (S)-N-hydroxy-2-(N-isopropoxybiphenyl-4-ylsulfonamido)-3-methylbutanamides. *Bioorg. Med. Chem.* **2006**, *14*, 4260–4276.
- (21) Li, W.; Zhang, J.; Wang, J.; Wang, W. Metal-Coupled Folding of Cys<sub>2</sub>His<sub>2</sub> Zinc-Finger. *J. Am. Chem. Soc.* **2008**, *130*, 892–900.
- (22) Falconi, M.; Altobelli, G.; Iovino, M. C.; Politi, V.; Desideri, A. Molecular dynamics simulation of Matrix Metalloproteinase 2: fluctuations and time evolution of recognition pockets. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 837–848.
- (23) Lu, D. S.; Voth, G. A. Molecular dynamics simulations of human carbonic anhydrase II: Insight into experimental results and the role of solvation. *Proteins* **1998**, *33*, 119–134.
- (24) Ryde, U. Molecular dynamics simulations of alcohol dehydrogenase with a four- or five-coordinate catalytic zinc ion. *Proteins: Struct., Funct. Genet.* **1995**, *21*, 40–56.
- (25) Seminario, J. M. Calculation of Intramolecular Force Fields from Second-Derivative Tensors. *Int. J. Quantum Chem.* **1996**, *60*, 59–65.
- (26) Bautista, E. J.; Seminario, J. M. Harmonic force field for glycine oligopeptides. *Int. J. Quantum Chem.* **2008**, *108*, 180–188.
- (27) Brandt, P.; Norrby, T.; Åkermark, B.; Norrby, P. O. Molecular mechanics (MM3\*) parameters for ruthenium(II)–polypyridyl complexes. *Inorg. Chem.* **1998**, *37*, 4120–4127.
- (28) Norrby, P.; Rasmussen, T.; Hallen, J.; Strassmen, T.; Houk, K. N. Rationalizing the Stereoselectivity of Osmium Tetroxide Asymmetric Dihydroxylations with Transition State Modeling Using Quantum Mechanics-Guided Molecular Mechanics. *J. Am. Chem. Soc.* **1999**, *121*, 10186–10192.
- (29) Nilsson, K.; Lecerof, D.; Sigfridsson, E.; Ryde, U. An automatic method to generate force-field parameters for heterocompounds. *Acta Crystallogr.* **2003**, *59*, 274–289.
- (30) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A second generation force field for the simulation of proteins, nucleic acids and organic molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
- (31) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. C.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; V Rotiz, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskora, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; AllLaham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Gordon, M. H.; Replogle, E. S.; Pople, J. A. *Gaussian 03*; Gaussian, Inc.: Pittsburgh, PA, 1998.
- (32) Case, D. A.; Pearlman, D. A.; Caldwell, J. W.; Cheatham, T.; Wang, J.; Ross, W. S.; Simmerling, C.; Darden, T.; Merz, K. M.; Stanton, R. V.; Cheng, A.; Vincent, J. J.; Crowley, M.; Tsui, V.; Gohlke, H.; Duan, Y.; Pitera, J.; Massova, I.; Seibel, G. L.; Singh, U. C.; Weiner, P.; Kollman, P. A. *AMBER 9*; University of California: San Francisco, CA, 2006.
- (33) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nuc. Acid. Res.* **2008**, *28*, 235–242.
- (34) Wachters, A. J. H. Gaussian basis set for molecular wavefunctions containing third-row atoms. *J. Chem. Phys.* **1970**, *52*, 1033–1036.
- (35) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. Farnesyltransferase—new insights into the zinc-coordination sphere paradigm: evidence for a carboxylate-shift mechanism. *Biophys. J.* **2005**, *88*, 483–494.
- (36) Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. Theoretical studies on farnesyl transferase: evidence for thioether product coordination to the active-site zinc sphere. *J. Comput. Chem.* **2007**, *28*, 1160–1168.
- (37) Tamames, B.; Sousa, S. F.; Tamames, J.; Fernandes, P. A.; Ramos, M. J. Analysis of zinc-ligand bond lengths in metalloproteins: trends and patterns. *Proteins* **2007**, *69*, 466–475.
- (38) Amin, E. A.; Truhlar, D. G. Zn Coordination Chemistry: Development of Benchmark Suites for Geometries, Dipole Moments, and Bond Dissociation Energies and Their Use To Test and Validate Density Functionals and Molecular Orbital Theory. *J. Chem. Theory Comput.* **2008**, *4*, 75–85.
- (39) 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. A. Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-Phase Quantum Mechanical Calculations. *J. Comput. Chem.* **2003**, *24*, 1999–2012.
- (40) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.
- (41) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graphics Modell.* **2006**, *25*, 247–260.
- (42) Hoops, S. C.; Anderson, K. W.; Merz, K. M. Force Field Design for Metalloproteins. *J. Am. Chem. Soc.* **1991**, *113*, 8262–8270.
- (43) Cieplak, P.; Cornell, W. D.; Bayly, C.; Kollman, P. A. Application of the multimolecule and multiconformational RESP methodology to biopolymers: Charge derivation for DNA, RNA, and proteins. *J. Comput. Chem.* **1995**, *16*, 1357–1377.
- (44) Jorgensen, W. L.; Chandrasekhar, J.; Madurs, J.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (45) Loncharich, R. J.; Brooks, B. R.; Pastor, R. W. Langevin dynamics of peptides: The frictional dependence of isomerization rates of N-actylananyl-N'-methyl amide. *Biopolymers* **1992**, *32*, 523–535.

- (46) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Nola, A. D.; Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (47) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. The Smoothed Particle Mesh Ewald method. *J. Chem. Phys.* **1995**, *103*, 8577–8593.
- (48) Nair, S. K.; Elbaum, D.; Christianson, D. W. Unexpected binding mode of the sulfonamide fluorophore 5-dimethylamino-1-naphthalene sulfonamide to human carbonic anhydrase II. Implications for the development of a zinc biosensor. *J. Biol. Chem.* **1996**, *271*, 1003–1007.
- (49) Ryckaert, J. P.; Ciccotti, G.; Berendsen, J. C. Numerical integration of the Cartesian equation of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Chem. Phys.* **1977**, *23*, 327–341.
- (50) Walker, R. C.; Crowley, M. F.; Case, D. A. The Implementation of a Fast and Accurate QM/MM Potential Method in AMBER. *J. Comput. Chem.* **2008**, *29*, 1019–1031.
- (51) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods Applications. *J. Comput. Chem.* **1989**, *10*, 221–264.
- (52) Holland, J. *Adaptation in Natural and Artificial Systems*; University of Michigan Press: Ann Arbor, MI, 1975.
- (53) Nakamoto, K.; Takemoto, J.; Chaw, T. L. Metal Isotope Effect on Raman Spectrum of  $[\text{Zn}(\text{NH}_3)_4]_2$  Crystals. *Appl. Spectrosc.* **1971**, *25*, 352–355.

CT900454Q

## Merging Implicit with Explicit Solvent Simulations: Polyethylene Glycol

Alok Juneja,<sup>†</sup> Jorge Numata,<sup>†</sup> Lennart Nilsson,<sup>‡</sup> and Ernst Walter Knapp<sup>\*,†</sup>

*Freie Universität Berlin, Institute of Chemistry & Biochemistry, Fabeckstr. 36a,  
D-14195 Berlin, Germany and Centre for Biosciences, Department of Biosciences and  
Nutrition, Karolinska Institutet, SE-141 83 Huddinge, Sweden*

Received February 6, 2010

**Abstract:** We constructed an accurate polyether force field for implicit solvent (IS) molecular dynamics (MD) simulations that matches local and global conformations of 1,2-dimethoxy-ethane (DME) and polyethylene glycol (PEG), respectively. To make appropriate force field adjustments for IS models of PEG, we used long-term MD simulation data of 1  $\mu$ s in explicit solvent (ES) based on the most recent CHARMM35 ether force field that includes adjustments for PEG in explicit water. In IS models, competition of attractive van der Waals (vdW) interactions between solute–solute and solute–solvent atom pairs is often not considered explicitly. As a consequence, the attractive vdW interactions between solute atom pairs that remain in IS models explicitly can yield equilibrium structures that are too compact. This behavior was observed in the present study comparing MD simulation data of the DME and PEG ES model with corresponding IS models that use generalized Born (GB) electrostatics combined with positive surface energy terms favoring compact structures. To regain balance of attractive vdW interactions for IS models, we considered the IS generalized Born with simple switching (GBSW) model in detail, where we turned off surface energy terms and reduced attractive vdW interactions to 90%, or we used alternatively even slightly negative surface energies. However, to obtain quantitatively the same local and global distributions of PEG conformers as in ES, we needed additional force field adjustments involving torsion potentials and 1–4 and 1–5 atom pair Coulomb interactions. This CHARMM ether force field, specifically optimized for IS simulation conditions, is equally valid for dimeric and polymeric ethylene glycol. To explore the conformational space of PEG with MD simulations, an IS GBSW model requires 2 orders of magnitude less CPU time than the corresponding ES model. About a factor of 5 of this gain in efficiency is due to the lack of solvent viscosity in IS models.

### Introduction

The behavior of molecules in solution depends fundamentally on the balance between solute–solute, solute–solvent, and solvent–solvent interactions. The type of solvent used determines the solvation and association behavior of molecules as well as their protonation and redox states in the

case of titratable or redox-active molecules and their conformations in the case of larger flexible molecules. The native structure of biological macromolecules and in particular of proteins is governed by interactions with water.<sup>1,2</sup> Biological macromolecules need to be under physiological conditions (i.e., to be in aqueous solution at specific temperatures, pH's, and ionic strengths) to adopt their native structure, which is a prerequisite to function appropriately.<sup>3–7</sup> Binding strengths of drugs, substrates, and inhibitors to proteins are strongly influenced by water.<sup>2,8–10</sup> Also, conformational dynamics and the function of proteins is solvent

\* Corresponding author phone: 0049-30-838-54387; fax: 0049-30-838-56921; e-mail: knapp@chemie.fu-berlin.de.

<sup>†</sup> Freie Universität Berlin.

<sup>‡</sup> Karolinska Institutet.

controlled.<sup>11–13</sup> On the other hand, the neighborhood of a protein surface influences also the behavior of the surrounding water. It restricts conformational variability of water, leading to reduced entropy,<sup>12,14,15</sup> and slows down its dynamics by about a factor of 2 to 4.<sup>16,17</sup> Thus, modeling and simulation of structure and dynamics of molecules and their interactions in solutions generally requires the consideration of interactions with solvent molecules in atomic detail. This is why, in approaches using molecular dynamics (MD) simulations, the considered molecules need to be embedded sufficiently well in a solvent environment.

In conventional MD simulations of solute–water systems, water molecules are described explicitly in atomic detail employing an atom-based molecular force field. Such MD simulations involve routinely a large number of atoms, where most of them belong to solvent molecules. Even in the best case, the computational costs of such molecular systems increase faster than linearly with the number of atoms, making MD simulations with explicit solvent (ES) models rather expensive.

Different procedures have been applied to reduce the CPU time required for MD simulations. One focus is to use implicit solvent (IS) models for water,<sup>18–26</sup> which diminish the number of atoms to be considered enormously. An additional efficiency bonus of IS models is the absence of solvent viscosity. As a consequence, the actually used elementary time step of MD simulations with an IS model can in reality correspond to a larger time interval, as has been found in IS generalized Born using molecular volume (GBMV) MD simulations when a low friction constant was used.<sup>27</sup> This would allow exploration of the conformational space of a solute molecule in less simulation time but bears the disadvantage that the dynamics are unrealistically fast. Water interacts with solute molecules in three ways: two direct and one indirect type of interaction. The former two are electrostatic<sup>28</sup> and van der Waals (vdW) interactions;<sup>12</sup> the latter is due to the hydrophobic effect.<sup>2,13,15,29–31</sup> The influence of hydrogen bonds (H-bonds) is generally accounted for by a suitable combination of electrostatic and vdW interactions.

Solvent modeled as explicit water screens Coulomb (and also vdW) interactions between solute atoms and competes with the direct solute–solute atom pair interactions (Coulomb as well as vdW). In an IS model, the solute–solvent electrostatic interactions is approximated using a dielectric continuum with large dielectric constant for the solvent ( $\epsilon = 80$  for water), while the solute atomic partial charges are embedded in a dielectric cavity of a small dielectric constant (generally  $\epsilon = 1$ ). In a simplified approach, the electrostatic boundary between a low and high dielectric medium in an IS model can be approximated by the surface separating the vdW solute volume (given by the merged volumes of solute atoms) from the solvent, which is used by GBSW.<sup>32–34</sup> More elaborate procedures use the molecular surface as in GBMV.<sup>35,36</sup> In fast analytical continuum treatment of solvation (FACTS),<sup>37</sup> still another procedure to effectively generate the interface between solute and solvent is used, which is guided by the principle to save CPU time.

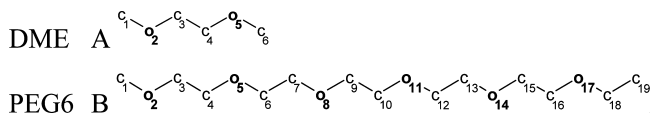
In dielectric continuum models, the solute atomic partial charges interact with virtual surface charges induced at the solute–solvent boundary. The corresponding electrostatic energies can be evaluated with the Poisson or for nonvanishing ionic strength with the Poisson–Boltzmann equation.<sup>22,38,39</sup> Since solving the Poisson equation at each time step of MD simulations slows down such simulations considerably, more approximate electrostatic models are used instead. These are, for instance, the generalized Born approximation (GB)<sup>40</sup> or even more simplifying approaches that use distance dependent dielectric constants and neutralized charged groups.<sup>41–44</sup> Although the continuum dielectric medium models explicit electrostatic solute–solvent interactions generally faithfully, it was occasionally observed that solute–solute H bonds and salt bridges are too persistent in the IS model with GB.

In addition to the electrostatic energy contributions, IS force fields contain a nonpolar energy term that accounts for the entropic costs for water to be in contact with the surface of a solute. In touch with the surface of a solute, a water molecule can no longer adopt as many different H-bond patterns as in bulk water. This goes along with a loss of entropy of these water molecules, the so-called hydrophobic effect. Consequently, water has positive contact energy with solute molecules forcing different solute molecules to aggregate and individual solute molecules to assume a compact conformation with a minimal surface. This nonpolar contribution to solvation free energy is generally assumed to be directly proportional to the solvent accessible surface area.<sup>45–47</sup> In MD simulations with the IS model, the hydrophobic effect is normally accounted for by an artificial energy term that is proportional to the solvent exposed solute surface<sup>45–47</sup> with a proportionality constant  $\gamma$  varying between 5 and 40 cal/(mol Å<sup>2</sup>).<sup>33,48–51</sup> There have been a number of efforts in recent times to advance IS models<sup>23,24,38,42,43</sup> to simulate dynamics of biomolecules efficiently and to come close to results of computationally more expensive conventional MD simulations with an ES model. However, some GB implementations have the tendency to become inefficient for large molecules.

IS models were also optimized to compute solvation energies efficiently.<sup>52–54</sup> There, it was found that electrostatic models combined with surface energy alone is not appropriate and must be supplemented by a volume term accounting for solvent cavity formation.<sup>52</sup> Here, we study the energetics of 1,2-dimethoxy-ethane (DME) and polyethylene glycol (PEG) conformers in an implicit solvent. Since their volumes practically do not vary with the conformers, these optimized IS models are not applicable in our case.

But there is still another artificial effect, which is generally present in IS models. In explicit water simulations (and also in real molecular systems), the attractive solute–solute and solute–solvent vdW interactions compete with each other and thus balance the composition of solute and solvent atoms in the neighborhood of solutes. In the absence of ES, only the attractive solute–solute vdW interactions are left, and as a result the neighborhood of solute molecules is unbalanced. These solute–solute interactions occur between atoms of different solute molecules (intermolecular) but also between atoms of the same solute molecule (intramolecular).





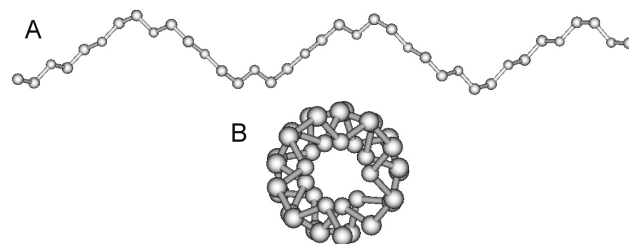
**Figure 1.** Schematic structure representations of DME (A) and PEG6 (B). Atoms are numbered to refer to specific numbering used in the text.

Removing the solvent molecules, the solute molecules interact differently. Since all atoms are subject to the mutual attractive vdW interactions, one observes for MD simulations in the absence of solvent, aggregation of different solute molecules and for flexible solute molecules also self-aggregation, leading to solute conformers which are too compact. The influence of solute–solvent vdW interactions on molecular solvation energies, ligand binding, and protein docking was discussed many times. However, to the best of our knowledge, it was first pointed out in ref 55 that this effect can also influence molecular conformations significantly. An IS force field including the effect from the solute–solvent vdW interactions appropriately by using a volume term was recently developed and applied to proteins.<sup>56</sup>

The hydrophobic effect<sup>13,15,29–31</sup> is based on the entropy content of the water structure, which assumes a maximum for undisturbed bulk water. While the hydrophobic effect has a physical basis, the mutual attraction of solute molecules in IS models, where attractive vdW interactions are not balanced, is an artifact. The hydrophobic effect causes an effective attraction between different solute molecules and in the case of flexible solute molecules also between atoms of the same solute molecule. It is properly considered in explicit water MD simulations<sup>16,57</sup> and is for instance the driving force of protein folding.<sup>6,7,12</sup> In implicit water simulations with vanishing surface energy, the effective attractive hydrophobic interaction between solute molecules is absent, while conversely the lack of balance in the attractive vdW interactions between solute–solute and solute–solvent contacts leads to an artificial attraction between solute molecules. Hence, under these conditions of an IS model, these two effects partially compensate each other.

Hence, in an IS model, proper tuning is required for the surface energy replacing the hydrophobic effect and the strength of the unbalanced vdW attractions. For similar reasons, it may be necessary to adjust Coulomb interactions for specific atom pairs of the solute to compensate for the absence of explicit H bonds with water molecules, which may have not been considered appropriately by the continuum electrostatic approach. Similar, more subtle effects can be due to a specific H-bond pattern between solute and water molecules, which may stabilize specific solute conformations and thus change the equilibrium distribution between them. As we will see, to account for all these effects appropriately, we need to adjust not only specific atom pair Coulomb interactions but also specific torsion potentials.

As a model system, we will study polyethylene glycol (PEG)<sup>58,59</sup> involving the ethylene glycol repeat unit ( $-\text{CH}_2-\text{CH}_2-\text{O}-$ ) (Figure 1). Crystallized and amorphous PEG adopt predominantly helix-like conformers (Figure



**Figure 2.** Side (A) and top (B) views of the PEG helix conformer composed of 12 monomer units in local TGT conformations comprising two helix turns. TGT is the most dominant local structure of PEG as well as of DME (see Figures S1 and S2 of the Supporting Information).

2).<sup>60,61</sup> They involve the most prevalent local structure TGT (see Figure S1, Supporting Information) found in 1,2-dimethoxy-ethane (DME) and PEG. *Ab initio* quantum chemical calculations of DME, which include solvent effects by continuum electrostatics<sup>62</sup> as well as MD simulations with umbrella sampling in explicit water,<sup>63</sup> indicate that DME prefers the TGT conformer, which is also evident from Raman,<sup>64</sup> IR,<sup>65</sup> and NMR studies.<sup>66</sup> The helical structure of PEG is stabilized by water bridges between nearest neighbor oxygen atoms in PEG. This has been reported by NMR,<sup>67</sup> the *ab initio* quantum chemical method,<sup>68</sup> and MD simulation<sup>69–71</sup> studies.

PEG is water-soluble, is nontoxic, degrades slowly by metabolic enzymes, and possesses low immunogenicity.<sup>72,73</sup> Therefore, it is widely used as an excipient in different pharmaceutical formulations, foods, and cosmetics.<sup>74,75</sup> PEG is also used as a precipitant agent for protein crystallization.<sup>76</sup> Being flexible and water-soluble, PEG can be used to create high osmotic pressure.<sup>77,78</sup> Coating gene therapy vectors with PEG reduces innate immune responses.<sup>79</sup> Self-assembled monolayers on a gold surface resist protein adsorption from aqueous solutions, if terminated with PEG.<sup>80–82</sup> PEG is of interest for the pharmaceutical industry to encapsulate drugs in nanoparticle structures or dendrimers made of PEG based polymeric material.<sup>75,83–86</sup> It can also be used to connect ligands serving as drugs to form dimeric or even multimeric ligands to enhance drug activity by the multivalent binding effect.<sup>87,88</sup> In multivalent ligand binding, the chain entropy of PEG plays a key role in understanding the effect quantitatively. So far, only a simple Gaussian chain model was used to describe the influence of the PEG linker on bivalent ligand binding.<sup>89,90</sup>

Because of the importance of PEG, it was the focus of many experimental<sup>81,91</sup> and computational studies.<sup>71</sup> Recently, the PEG force field parameters<sup>92,93</sup> for explicit water MD simulations were optimized in CHARMM.<sup>94,95</sup> The present study on IS generalized Born with simple switching (GBSW)<sup>32–34</sup> models uses results of MD simulations with ES based on this improved force field as a reference. Here, we like to demonstrate how large deviations in local and global conformational features of PEG can be, if one uses state of the art IS models like GBSW<sup>32–34</sup> for a flexible molecule, and what one has to change to obtain a faithful force field for IS GBSW MD simulations. We also try to show for DME and PEG to what extent two more GB based IS models (GBMV<sup>35,36</sup> and FACTS<sup>37</sup>) implemented in

CHARMM<sup>94,95</sup> may yield MD simulation data deviating from data obtained with corresponding ES models.

## Methods

**General Considerations.** In the recent years, much effort was placed in the development of theoretical models explaining the conformational preference of 1,2-dimethoxy-ethane (DME).<sup>58</sup> For this purpose, specialized classical force field parameters for MD simulations of DME and PEG were developed<sup>96,97</sup> and applied,<sup>98,99</sup> indicating the existence of specific DME conformers in the liquid phase. Recently, the CHARMM force field has been complemented with parameters for ether compounds,<sup>92</sup> which were further adjusted in CHARMM35<sup>93</sup> using DME Raman spectra.<sup>64</sup>

The DME structure can be characterized by three dihedral angles defining the local arrangement of a set of four covalently connected atoms (C–O–C–C, O–C–C–O, and C–C–O–C). Each of the torsion angles can be in one of the three possible torsion angle intervals, which are centered around 180°, called the trans (T) conformer, or around +60° or –60° corresponding to the two possible gauche (G, G′) conformers. In MD simulations with the ES model, DME adopts a mixture of mainly four conformers, which ordered with decreasing population are TGT, TGG′, TGG, and TTT (see Figures S1 and S2 of the Supporting Information and the insert of Figure 4A). These DME conformers have been observed in various solvents like water,<sup>100</sup> methanol, and carbon tetrachloride.<sup>101</sup> Among these four principal conformers, it has been noticed that TGT has the lowest free energy. In experiments,<sup>100,101</sup> one observes an interesting inverse correlation between the gauche C–C bond and C–O bond populations, which is prevalent in DME and PEG chains and thus corroborates the choice of DME as a model compound to adjust the PEG force field.<sup>93</sup>

Inspired by the Born equation to compute solvation energies of ions,<sup>102</sup> a number of generalized Born (GB) models<sup>33,36,37,40,49,50,103–105</sup> have been developed in the past years. In GB models, the effective nonpolar interactions related to the hydrophobic effect are generally proportional to the solvent accessible surface area (SA).<sup>106</sup> In the actual CHARMM force field, version 35 (CHARMM35),<sup>93</sup> the GBSW<sup>32–34</sup> module for IS MD simulations uses as a default value the proportionality constant  $\gamma = 30 \text{ cal}/(\text{mol } \text{Å}^2)$  for the surface energy term.

To optimize an IS force field for PEG, we considered the GBSW<sup>32–34</sup> model in detail to evaluate the electrostatic interactions as implemented in CHARMM combined with the ES ether force field parametrization of CHARMM35.<sup>93</sup> To probe the generalities of our results found with the GBSW<sup>32–34</sup> model, we also explored two more GB models available in CHARMM (GBMV<sup>35,36</sup> and FACTS<sup>37</sup>). Although we did not consider the GB models in AMBER<sup>107,108</sup> explicitly, their parametrization is qualitatively similar to GB models considered in the present study, such that we expect analogue behavior.

Henceforth, we refer to MD simulations using CHARMM35 with the TIP3P water model<sup>109</sup> as an explicit solvent (ES) model and in the absence of a solvent as described above

with default CHARMM35 parametrization as the IS GBSW model with a positive surface energy. The different IS GBSW models considered in the present study are labeled by additional information where they deviate from the default CHARMM35 parameter settings. For the other two GB models (GBMV and FACTS) considered here, we varied the surface energy only. The computations based on the IS GBMV model use the CHARMM35 force field with the recent adjustments for PEG,<sup>93</sup> while the IS FACTS model is based on the CHARMM22<sup>95</sup> force field.

**Scaling CHARMM Interactions for Implicit Solvent Simulations.** The polyether torsion potentials  $V_{\text{O–C–C–O}}(\phi)$  and  $V_{\text{C–O–C–C}}(\phi)$  describing the rotation barrier for the C–C and O–C bonds, respectively, were for explicit solvent simulations optimized yielding in units of kcal/mol

$$V_{\text{O–C–C–O}}(\phi) = 0.59(1 + \cos(\phi - \pi)) + 1.16(1 + \cos(2\phi)) \quad (1a)$$

$$V_{\text{C–O–C–C}}(\phi) = 0.57(1 + \cos(\phi)) + 0.29(1 + \cos(2\phi)) + 0.43(1 + \cos(3\phi)) \quad (1b)$$

In order to match the torsion angle distributions between ES and IS GBSW model simulations, the parameter of the first term in  $V_{\text{O–C–C–O}}(\phi)$  was increased from 0.59 to 1.09, while the parameter of the second term in  $V_{\text{C–O–C–C}}(\phi)$  was reduced from 0.29 to 0.20.

Furthermore, we varied the surface energy term described by the surface tension coefficient  $\gamma$ ,<sup>21,110</sup> whose value is proposed to be 30 cal/(mol Å<sup>2</sup>)<sup>32,33</sup> in the CHARMM35 IS force field. We also varied the attractive  $r^{-6}$  term of the LJ atom-pair interactions. Fine tuning of the energy function of the IS GBSW model required also a change in the Coulomb interactions of the 1–4 atom pairs and of the 1–5 O–H (e.g., O<sub>2</sub>–H<sub>6</sub> and O<sub>5</sub>–H<sub>1</sub> where the subscript at H refer to the carbon atom to which the hydrogen atom is attached to, see Figure 1) atom pairs.

**MD Simulation Protocols.** The program CHARMM was used to prepare the ES MD simulation setup with the CHARMM35 ether force field.<sup>93</sup> The TIP3P model<sup>109,111</sup> was used for water. The long-term MD simulations were performed by NAMD.<sup>112</sup> These were 0.1  $\mu\text{s}$  for DME with 199 TIP3P water and 1  $\mu\text{s}$  for PEG6 with 1477 TIP3P water. All parameters and conditions for NAMD were the same as for CHARMM. The temperature was controlled by Langevin thermostat with friction coefficient  $\beta = 2 \text{ ps}^{-1}$ , and electrostatic interactions were evaluated using the particle mesh Ewald method.<sup>113</sup> The first 1 ns was discarded to allow for equilibration. Atomic coordinates were saved every 0.2 ps time step. More details on the ES simulation conditions are given in the Supporting Information.

For the three different IS models, MD simulations were carried out with CHARMM35 using the modules GBSW,<sup>32–34,104</sup> GBMV,<sup>35,36</sup> and FACTS.<sup>37</sup> A canonical NVT ensemble was used maintaining the temperature at 300 K with the Nose–Hoover thermostat.<sup>115,116</sup> For GBSW and GBMV, the ether force field<sup>93</sup> was used. To define the electrostatic boundary, GBSW used the vdW surface with optimized atomic Born radii,<sup>114</sup> GBMV used the molecular surface, while FACTS used the solvent accessible surface

area approach. For DME and PEG6 MD, simulations of 100 and 400 ns (200 ns for GBMV) were performed, respectively. Atomic coordinates are saved every 0.2 ps. More details on the simulation conditions are given in the Supporting Information.

**Comparison of MD Simulation Data of Explicit and Implicit Solvent Models.** To compare the behavior of the DME and PEG force fields using the ES and the IS GBSW models, we considered specific atom pair distance distributions, which were evaluated with standard histogram techniques using the conformers from MD simulation data. The torsion potentials of the IS GBSW models for DME and PEG were optimized by comparing the corresponding torsion potentials. These were obtained by first evaluating the probability distribution of the torsion angles from the conformers of the MD simulation data by a histogram method. These probability distributions were then transformed to free energies by taking the negative logarithm of these probabilities and multiplying them by  $k_B T$  with  $T = 300$  K.

How efficient MD simulations explore the space of PEG conformers is found out by observing how the end-to-end atom pair distance distributions  $g^{(t)}(x)$  evolve with the total time span  $t$  used for the ensemble averages. A suitable quantity to measure how fast the limit distribution  $g^{(\infty)}(x)$  of the ensemble is approached with the MD simulation time is the integral of the square deviation

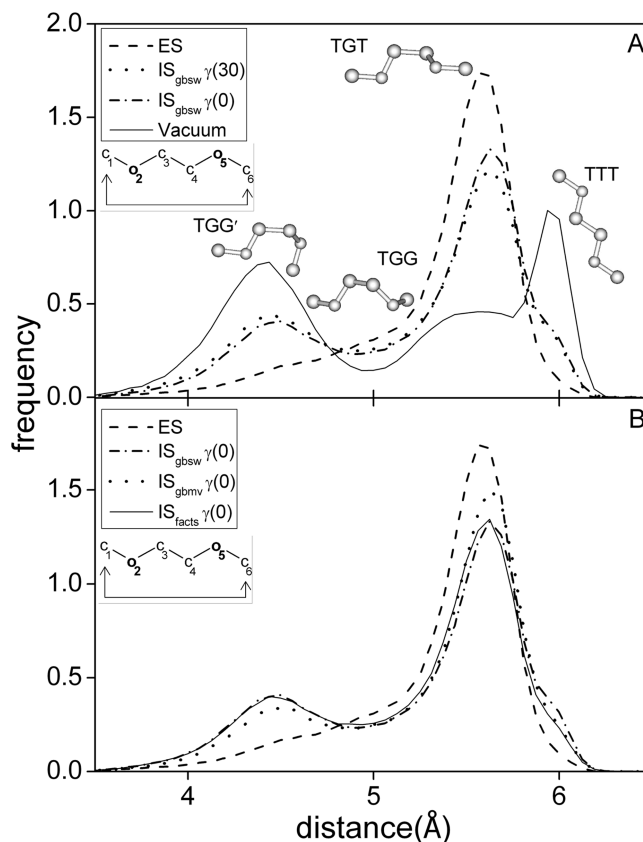
$$[\Delta g^{(t)}]^2 = \int [g^{(t)}(x) - g^{(\infty)}(x)]^2 dx \quad (2)$$

Alternatively, one can monitor the end-to-end distance autocorrelation function of the time dependent distance  $d(t)$

$$c(\Delta t) = \frac{\langle [d(t) - \bar{d}][d(t + \Delta t) - \bar{d}] \rangle_t}{\langle [d(t) - \bar{d}]^2 \rangle_t} \quad (3)$$

## Results and Discussion

**General Considerations.** A general overview of typical results from state of the art MD simulations applied to DME can be found in Figure 3, where the end-to-end distance distribution is displayed for simulations under conditions of vacuum, ES, and IS models, i.e., GBSW,<sup>32,33</sup> GBMV,<sup>35,36</sup> and FACTS.<sup>37</sup> It is not surprising that differences of DME conformations between explicit water and a vacuum are enormous (see dashed and solid lines in Figure 3A, respectively). But, conformational distributions of DME obtained with IS simulation conditions can still differ considerably from results of ES MD simulations. In particular, we observed that the most prominent TGT conformer is for all values of surface tension parameter  $\gamma$  less populated with IS than with ES simulation conditions (see Figure 3 and Figure S3, Supporting Information). With the IS model having positive surface energy, DME conformers are generally too compact due to the positive surface energy term and the unbalanced vdW attractive  $r^{-6}$  terms that preferentially populate the TGG' and TGG conformers (see dotted line in Figure 3A and dotted lines in Figure S3, Supporting Information). Although the DME conformer distribution obtained with the IS model with positive surface energy



**Figure 3.** Conformational distribution of DME monitored by the  $C_1-C_6$  end-to-end distance obtained from MD simulations under different force field conditions. (A) Comparison of ES results with data based on a vacuum and the IS GBSW model: vacuum conditions (solid line); explicit water (dashed line); IS GBSW with vanishing surface energy (dashed-dotted line); IS GBSW model with positive surface energy [ $\gamma = 30$  cal/(mol  $\text{\AA}^2$ )] (dotted line). (B) Comparison of ES results with three different IS models, all with a vanishing surface energy term: explicit water (dashed line); IS GBSW<sup>32-34</sup> (dashed-dotted line); IS GBMV<sup>35,36</sup> (dotted line); IS FACTS<sup>37</sup> (solid line).

shows large deviations relative to results from the ES model, the TGT conformer corresponding to the helix conformation of PEG (see Figure 2) remains the most populated one. Hence, the electrostatic screening effect due to an unspecific H-bond pattern of DME-water interactions, which is modeled by GB electrostatics, approximates qualitatively the most prominent features of the DME conformer distribution. The DME conformer distributions obtained with the IS GBSW model show compared to commonly used positive surface energies [ $\gamma = 30$  cal/(mol  $\text{\AA}^2$ )] relatively small variations for moderate negative surface energies [ $\gamma = -15$  cal/(mol  $\text{\AA}^2$ )] (see Figure S3A, Supporting Information). Under the same conditions, the other two IS models available in CHARMM (GBMV<sup>35,36</sup> and FACTS<sup>37</sup>) show comparatively larger variations (see Figure S3B and C of the Supporting Information).

In the following, we like to demonstrate what is responsible for the deficiencies of the IS GBSW model and how one can bridge the gap to reach better agreement with ES simulation data for DME and PEG. To optimize the energy function for the IS GBSW model, we first considered DME. For this purpose, we studied how the MD simulation data



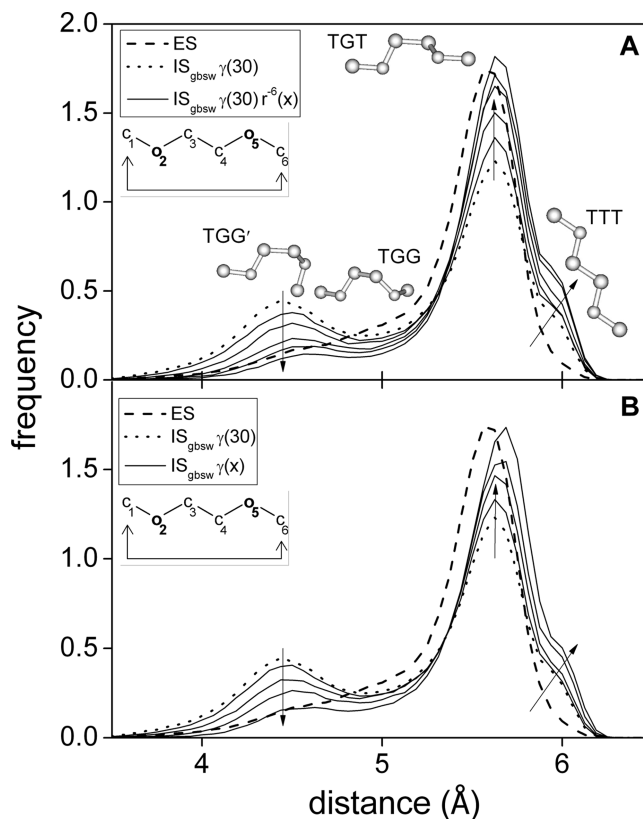
for the IS GBSW DME model depend on the strength of the vdW attraction and the surface energy term. Next, we investigated how the IS GBSW DME model varies with the strength of specific Coulomb interactions. We optimized these above-mentioned interactions before we started to fine-tune the energy function with respect to the torsion potentials. Finally, we investigated whether the DME optimized energy function can be transferred to PEG.

As we will see later, a more accurate DME conformer distribution is obtained with an IS GBSW model with slightly reduced 1–4 Coulomb interactions, indicating a subtle but relevant contribution due to the specific H-bond pattern of water in the neighborhood of DME and PEG. This qualitatively correlates with earlier studies on the importance of the H-bond pattern between PEG and water that stabilizes the helical structure of PEG.<sup>67–71</sup>

**Comparison of MD Simulation Data of DME with Explicit and Implicit Water Models and the Role of vdW Attraction and Surface Energy.** We first like to compare MD simulation data of DME (see Figure 1), obtained with the ES and IS GBSW model with positive surface energy. Searching for a data representation, which shows the differences between the conventional ES and IS GBSW models most clearly, we found that this is the case for the end-to-end distance distributions of the DME atoms C<sub>1</sub> and C<sub>6</sub> (Figure 1). The dashed line in Figure 4 shows the distance distributions for the ES model and the dotted line for the IS GBSW model. One can clearly observe that, with the IS GBSW model, the compact DME structure at 4.5 Å is considerably more populated at the expense of the more extended structure at about 5.5 Å (see Figure 4A, dotted line compared to dashed line and schematic DME structures therein). Furthermore, in the IS GBSW model, a shoulder appears at 6.0 Å end-to-end distances, which is absent in the ES model. The enhanced occurrences of the compact structures in the IS GBSW model are due to unbalanced vdW interactions, as we will see.

One can qualitatively correct for the enhanced occurrence of compact structures by decreasing the attractive wing of the LJ interaction (except for the 1–4 atom pair interaction, which is part of the corresponding torsion potential and therefore should not be changed), or by decreasing the surface energy term into the negative regime as shown in the top and bottom parts of Figure 4, respectively. To account for unbalanced vdW interactions, negative surface energies were also considered to characterize the energetics of protein mutants.<sup>23</sup> However, negative surface energies were occasionally also used for IS GBSW models to account for dielectric screening, in the absence of more expensive electrostatic approaches like GB or PB.<sup>24</sup>

With extreme values of negative surface tension [ $\gamma = -200$  cal/(mol Å<sup>2</sup>)], the effect from the surface energy term saturates (Figure 4B). But, even with such extreme corrections, the end-to-end distance distribution of DME obtained with the IS GBSW model shows still marked deviations relative to the reference distribution obtained for conventional ES MD simulations (dashed line in Figure 4B): The side maximum at 4.5 Å remains, while it appears in the ES MD simulation as a shoulder only. Furthermore, a significant

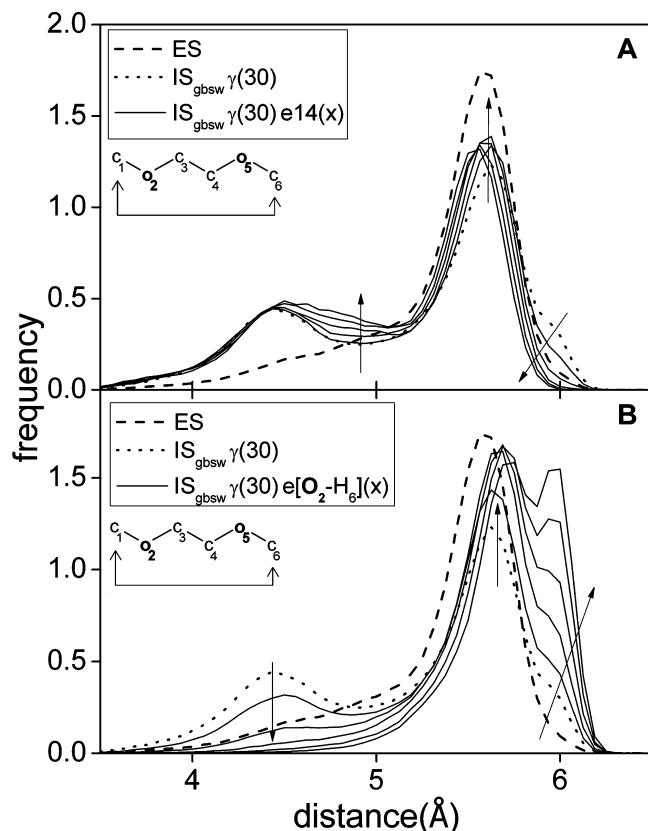


**Figure 4.** Role of attractive vdW and surface energy. The end-to-end distance distributions of DME obtained from MD simulation is shown for the ES (dashed line) and IS GBSW models with positive surface energy (dotted line). (A) The  $r^{-6}$  vdW attraction (except the 1–4 atom pair interaction) is scaled down by factors of 0.9, 0.7, 0.5, 0.3, and 0.1 (solid lines). The four most dominant DME structures are shown for the maxima at 4.5 and 5.5 Å and the shoulder at 6.0 Å, corresponding to structures where subsequent three torsion angles (C–O–C–C, O–C–C–O, C–O–C–C) correspond to TGT, TGG', TGG, and TTT, respectively. (B) The surface tension coefficient  $\gamma$  is varied from 0, –50, –100, and –200, cal/(mol Å<sup>2</sup>) (solid lines). Note that for these exceedingly negative surface energies, variations of DME conformers can be observed also with the IS GBSW model favoring conformers with larger surfaces. The arrows indicate the direction of the changes in the distribution occurring by decreasing either the vdW attraction (A) or the surface tension coefficient (B).

shoulder appears at 6.0 Å, referring to the most extended DME structure (i.e., the TTT conformer), and the main maximum is shifted to larger distances. A similar behavior in the end-to-end distance distribution of DME is observed, if the attractive  $r^{-6}$  term of the vdW interaction is lowered from 1.0 by up to a factor of 0.1 (see Figure 4A).

**Implicit Water Models of DME and the Role of 1–4 and 1–5 Coulomb Atom Pair Interactions.** Exploring different options to improve the agreement of MD simulation data of DME, it turned out to be useful to first optimize the nonbonded interactions. Although the GB approach accounts for electrostatic contributions of solute–solvent interactions, there may be deficiencies, which need to be corrected. The strongest electrostatic interactions of DME and PEG involve ether oxygen pairs and ether oxygen with nonpolar hydrogens. Hence, we studied the dependence of the end-to-end





**Figure 5.** Role of 1–4 and 1–5 O–H atom pair interactions for the end-to-end distance distributions of DME obtained from MD simulation. In A and B, the dashed line refers to the ES and the dotted line to the IS GBSW model with positive surface energy. (A) The 1–4 atom pair Coulomb interaction is scaled down by factors of 0.9, 0.7, 0.5, 0.3, and 0.1 (solid lines). (B) The 1–5 oxygen–hydrogen (O<sub>2</sub>–H<sub>6</sub>, O<sub>5</sub>–H<sub>1</sub>) Coulomb interaction is scaled down by factors of 0.9, 0.7, 0.5, 0.3, and 0.1 (solid lines). The arrows indicate the direction of the changes in the distribution occurring by decreasing either the 1–4 atom pair (A) or the oxygen–hydrogen (B) electrostatic interaction.

distance distribution of DME on the Coulomb interactions of all 1–4 and oxygen–hydrogen 1–5 atom pairs (O–H; e.g., O<sub>2</sub>–H<sub>6</sub> and O<sub>5</sub>–H<sub>1</sub>, where H<sub>i</sub> is attached to C<sub>i</sub>), shown in Figure 5 (solid lines), top and bottom parts, respectively. We suspected that the direct Coulomb interactions for these atom pairs may be overestimated. Therefore, they were scaled down by factors of 0.9, 0.7, 0.5, 0.3, and 0.1 (Figure 5).

Scaling down the 1–4 atom pair Coulomb interactions (Figure 5A) removes the shoulder at 6.0 Å and simultaneously can partially fill the valley between the side (at 4.5 Å) and main maximum (at 5.5 Å) of the end-to-end distance distribution. The former is mainly due to the 1–4 O–O electrostatic repulsion. Furthermore, while the height of the main maximum remains essentially invariant, its position is shifted slightly to lower distances. All these changes are useful to correct deficiencies that remained or appeared after diminishing vdW attraction or surface energy as discussed in the preceding subparagraph.

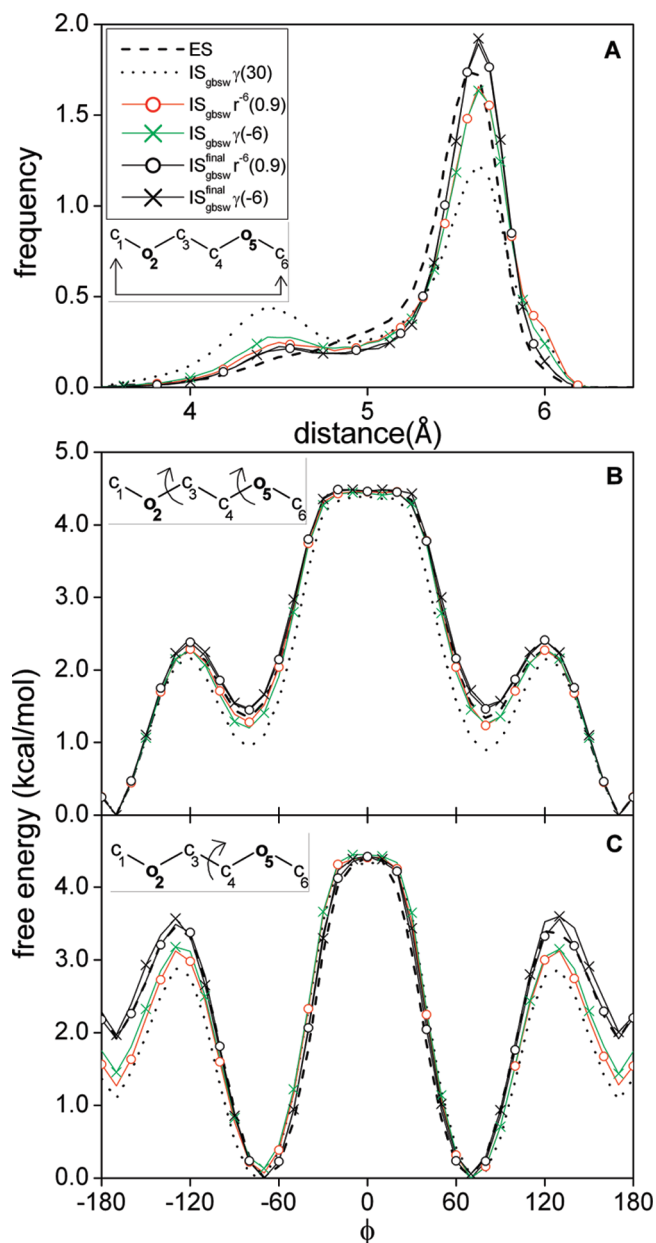
The 1–5 O–H interaction governs the strength of the weak H bond between O<sub>2</sub> (O<sub>5</sub>) and the methyl group at position 6 (1). We also suspected that this direct 1–5 O–H

interaction may be weaker in explicit solvent. Scaling down this interaction converts the side maximum at 4.5 Å to a shoulder similar to the ES model but of lower intensity. Simultaneously, the shoulder at 6.0 Å corresponding to all-trans stretched conformation (see Figure 5B) grows considerably, and the main maximum shifts to larger distances. Although the latter is unfavorable, in combination with the scaling of the 1–4 Coulomb interaction, it can be useful.

**Implicit Water Models of DME: Fine-Tuning of Energy Function.** The next step is to optimize the above-discussed four different nonbonded energy terms, which are (1) the attractive vdW term, (2) the surface energy, (3) the 1–4 atom pair Coulomb energy, and (4) the 1–5 O–H Coulomb energy. This was done in a tedious manual optimization procedure, where we first monitored the end-to-end distance distribution of DME (see Figure 6A). In doing so, we found two possible nearly equivalent solutions, which have in common that the interactions 3 and 4 are reduced to 90%. One option is with vanishing surface energy ( $\gamma = 0$ ) where the vdW attractive  $r^{-6}$  term is reduced to 90% (red lines with open circles  $\circ$  in Figure 6). The second option is to assume a slightly repulsive surface energy term [ $\gamma = -6$  cal/(mol Å<sup>2</sup>)] with no change in the vdW attractive interaction (green lines with  $\times$  in Figure 6). The corresponding end-to-end distance distributions are already close to the results obtained with the ES model (dashed line in Figure 6) that is used as a reference, but there is still room for improvement.

Alternatively to the reduction of specific Coulomb interactions as explained above (see also Table 1) to 90% of their original values, we lowered the Born radius of ether oxygen from 1.80 Å to 1.52 Å. The latter is the Born radius of backbone oxygen used in CHARMM.<sup>94,95</sup> Due to the stronger solvation of such oxygens, the corresponding explicit solute–solute interactions are effectively weakened, yielding better agreement with ES simulation data (see Figure S6 of the Supporting Information).

Optimization of the torsion potentials was done in the next step. For this purpose, we observed the free energy profiles of the torsion angles as derived from MD simulation data. We show the results of MD simulations for the torsion potentials involving rotations around the two C–O bonds and the C–C bond of DME in parts B and C of Figure 6, respectively. Most relevant is the C–C bond rotation, which shows still larger deviations from the MD simulation data obtained with the ES model (compare the dashed line with the green and red line in Figure 6C). Optimizing the parameters of the torsion potentials to the values given in Table 1, we obtain the black solid lines in Figure 6 marked with symbols  $\circ$  ( $\times$ ) referring to vanishing surface energy ( $\gamma = 0$ ) and vdW attraction at 90% (negative surface energy [ $\gamma = -6$  cal/(mol Å<sup>2</sup>)] and unchanged vdW interaction). With this choice of the torsion potentials, we are essentially closing the gap appearing in the effective C–C bond torsion potential (Figure 6C) between the results of the ES reference model and the IS GBSW models. Slight improvements can also be observed for C–O torsion angle distribution and the end-to-end distance distribution (Figure 6A,B). The 1–5 O–C (O<sub>2</sub>–C<sub>6</sub> and C<sub>1</sub>–O<sub>5</sub>) and the 1–4 O–O (O<sub>2</sub>–O<sub>5</sub>) and



**Figure 6.** DME simulation data. Dashed (dotted) lines are the results based in the ES (IS GBSW with positive surface energy) model. All other data have in common that the 1–4 and 1–5 O–H atom pair Coulomb interactions are reduced to 90%. Red (green) lines with ‘o’ (‘x’) symbols display MD simulation data where the surface energy vanishes ( $\gamma = 0$ ) and the attractive vdW interaction is reduced to 90% (the surface energy is negative [ $\gamma = -6$  cal/(mol Å<sup>2</sup>)] and the vdW interaction unchanged). The black solid lines show the final optimized MD simulation data including also the torsion potential corrections as described in the method section and Table 1. (A) End-to-end (C<sub>1</sub>–C<sub>6</sub>) distance distribution of DME. (B) Free energy profiles averaged over the two C–O–C–C torsion angles (C<sub>1</sub>–O<sub>2</sub>–C<sub>3</sub>–C<sub>4</sub> and C<sub>3</sub>–C<sub>4</sub>–O<sub>5</sub>–C<sub>6</sub>). (C) Free energy profile of the O–C–C–O torsion angle (O<sub>2</sub>–C<sub>3</sub>–C<sub>4</sub>–O<sub>5</sub>). DME simulation data obtained by using the GB model only (IS with vanishing surface energy) are shown in Figures S4 and S5 of the Supporting Information.

distance distributions of DME are shown in Figure S4B and C of the Supporting Information, respectively. They also agree well with the MD simulation data of the ES model.

The force field parameters for the ES and IS GBSW model with positive surface energy and the two optimized IS GBSW models of DME are collected in Table 1. In the Supporting Information, we compare the populations of DME conformers obtained with the bare GB and the IS GBSW model with positive surface energy models for the atom pair distances (Figure S4, Supporting Information) and the torsion potentials (Figure S5, Supporting Information). These populations exhibit only moderate variations, since for moderate-sized surface tensions, the surface practically does not vary between the four most populated DME conformers (TGT, TGG′, TGG, TTT; see  $G_{\text{surface}}$  values in Figure S1 of the Supporting Information).

**MD Simulation of PEG in Implicit Solvent with the Fine-Tuned Energy Functions.** We now explore whether the IS GBSW model developed for DME can be transferred to PEG. For this purpose, we compare MD simulation data of PEG6 (involving 6 monomer units) based on the ES model with data of PEG6 obtained with the two IS GBSW models that were optimized for DME.

We first analyze the local conformers of PEG6. The C<sub>1</sub>–C<sub>6</sub> atom pair distance distributions of PEG6 (Figure S7A of the Supporting Information) are virtually identical to the corresponding end-to-end atom pair distribution of DME (compare Figure 6A). With the two optimized force fields, one obtains for PEG6 practically the same agreement with the data based on the ES model as for DME. The same is true for the distance distributions of 1–5 O–C (O<sub>2</sub>–C<sub>6</sub>) (Figure S7B of the Supporting Information) and 1–4 O–O (O<sub>2</sub>–O<sub>5</sub>) atom pairs (Figure S7C of the Supporting Information), where the corresponding distance distributions for DME are shown in Figure S4B and C of the Supporting Information. The torsion potentials obtained for PEG6 that correspond to the DME data shown in Figure 6B and C are displayed in Figure S8A and B of the Supporting Information. The corresponding distance distributions of PEG6 for more distant O–O atom pairs are shown in Figure S10 of the Supporting Information.

The global behavior of the PEG6 conformers is probed by the end-to-end C<sub>1</sub>–C<sub>19</sub> atom pair distance distribution. Agreement with the global behavior of conformers obtained with the ES model is much more demanding for the larger PEG6 chain molecules than for the small DME. Nevertheless, the agreement of the optimized IS GBSW models (Figure 7A, solid lines with O (x) symbols) is fairly good, while the IS GBSW model with positive surface energy fails (Figure 7A, dotted line).

To probe the generalities of our results found for PEG6 with the GBSW<sup>32–34</sup> model, we also explored the other two GB models available in CHARMM, i.e., GBMV<sup>35,36</sup> and FACTS,<sup>37</sup> using positive [ $\gamma = 30$  cal/(mol Å<sup>2</sup>)] and vanishing surface energies [ $\gamma = 0$  cal/(mol Å<sup>2</sup>)]. Under the same conditions, the GBMV model shows large variations for positive and vanishing surface energies. The latter is in good agreement with the ES results (Figure 7B). For DME (Figure S3B of Supporting Information), the corresponding structural variations obtained with GBMV are comparatively moderate.

On the other hand, the PEG6 end-to-end distance distribution obtained with FACTS shows only small variations (even smaller than they appear for DME; see Figure S3C of

**Table 1.** Force Field Parameters Used for DME with Explicit Solvent (ES) and Implicit Solvent (IS) GBSW Models

force field term	CHARMM35 force field		adjusted CHARMM35 force field <sup>a</sup>	
	ES	IS <sup>b</sup>	IS <sup>final</sup> <sub>r<sup>-6</sup></sub> (0.9) <sup>b</sup>	IS <sup>final</sup> <sub>γ(-6)</sub> <sup>b</sup>
C–O–C–C eq (1b)	0.29/2/0	0.29/2/0	0.20/2/0	0.20/2/0
O–C–C–O eq (1a)	0.59/1/180	0.59/1/180	1.09/1/180	1.09/1/180
vdW (no14)	1.0	1.0	0.9	1.0
e14fac	1.0	1.0	0.9	0.9
elec O(1)–H(6)	1.0	1.0	0.9	0.9
γ [cal/mol Å <sup>2</sup> ]	N/A	30	0	-6

<sup>a</sup> The two adjusted CHARMM35<sup>93</sup> ether force fields both involve down-scaling of 1–4 all atom and 1–5 O–H atom pair Coulomb interactions to 90% and optimized parameters of torsion potentials (see eqs 1a and 1b). The IS GBSW models IS<sup>final</sup><sub>r<sup>-6</sup></sub>(0.9) and IS<sup>final</sup><sub>γ(-6)</sub> force fields refer to the cases of (i) vanishing surface energy (γ = 0) and 90% of attractive vdW interaction and (ii) negative surface energy [γ = -6 cal/(mol Å<sup>2</sup>)] and vdW interaction unchanged. <sup>b</sup> IS GBSW model.<sup>32–34</sup>

Supporting Information) with positive and vanishing surface energy terms and good agreement with the ES simulation data in both cases (Figure 7C). The insensitivity of FACTS on the variation of the surface energy in the end-to-end distance distribution of PEG6 as compared to a strong dependence found in the IS models of GBSW and GBMV may be related to the different way FACTS evaluates the effective molecular surface separating a low from a high dielectric medium. Naturally, the distributions of the local conformations of DME and consequently also of PEG do not agree so well for GBMV and FACTS, since they are not specifically optimized for DME and PEG, as it was done for the IS GBSW model in the present study.

While the maximum of the end-to-end distance distribution is at 13.5 Å for the ES model, it is only at 7 Å for the IS GBSW model with positive surface energy. The most probable distance of PEG6 in the ES model is considerably smaller than the end-to-end distance in the ideal TTT and TGT helix conformations, which are 24 Å and 18.5 Å, respectively. The distance of the energy minimized PEG6 helix conformer remains with about 17.2 Å for the IS GBSW model with positive surface energy and the optimized IS GBSW model close to value of the ideal helix structure of PEG. In addition to end-to-end distance distribution, the radius of gyration is also measured to assess the global structure of PEG6 and is shown in Figure S11 of the Supporting Information.

For the IS GBSW model with positive surface energy, the PEG6 conformers are generally much more compact due to the attractive surface energy term and the unbalanced vdW attractive  $r^{-6}$  terms that both predominantly populate the TGG' and TGG conformers (see Figure 7A, dotted line). However, also the attractive Coulomb interactions of the 1–5 O–H atom pairs that preferentially populate the TGG' state of DME (see insert in Figure 4A) may contribute to the more compact PEG6 conformers of the IS GBSW model with positive surface energy. One may suspect that it may also be favorable to reduce the Coulomb interactions for all atom pairs of the type 1–5 O–H to 90%. We tried it but obtained PEG6 conformers that were much too extended (results not shown). Hence, the best agreement with the ES model of PEG is obtained, if the force field optimized for the DME IS GBSW model is directly transferred to PEG.

Interestingly, the bare GB model without surface energy term applied to PEG6 yields an end-to-end distance distribution very close to the results obtained with the ES model

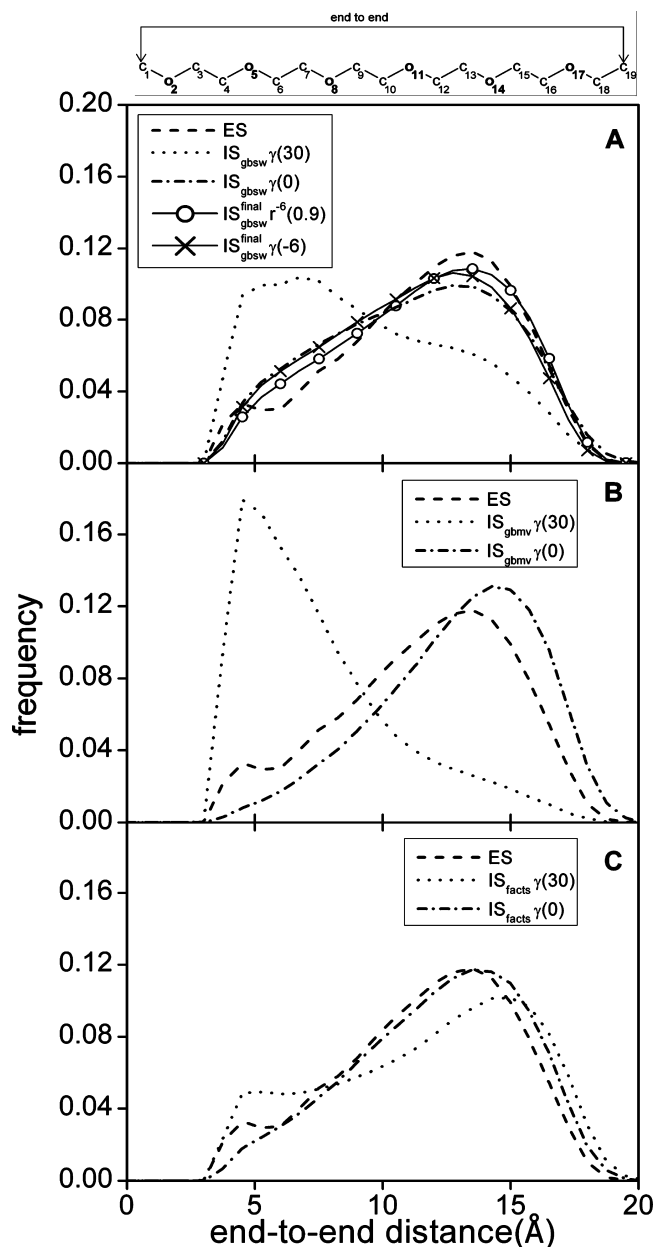
(see Figure 7A, dashed-dotted line). However, the local PEG conformations obtained with the bare GB model differ from the ES model for both PEG6 and DME in the same way (compare Figures S7 and S8 as well as Figures S4 and S5, Supporting Information). Using small or nearly vanishing surface energies agrees with recent applications to compute solvation energies.<sup>48–50,52,117</sup>

**Exploring the Conformational Space of PEG with MD Simulations.** We evaluated the end-to-end distance autocorrelation functions of PEG6 according to eq 3, which exhibits a stretched exponential decay behavior  $\exp(-t^\beta)$  with  $\beta = 2/3$  (Figure S12B of the Supporting Information) for the ES model, while for the IS GBSW models the decay behavior obeys a power law  $t^{-\alpha}$  with  $\alpha = 1/2$  (Figure S12A of the Supporting Information). A stretched exponential decay behavior with  $\beta = 1/2$  was for instance also found for the Rouse polymer model in viscous media, which is typical for dynamics governed by defect diffusion.<sup>118</sup> Evidently the IS GBSW models show long-term dynamics, which differs from the dynamics of the ES model. The end-to-end distance autocorrelation function of PEG6 decays by 2 orders of magnitude in about 300 ps for the IS GBSW model and in about 1 ns for the ES model.

It is interesting to compare how efficient ES and IS GBSW models of PEG explore the conformational space with time. We expected that the behavior of the two models will differ, since in the absence of explicit solvent the dynamics of IS GBSW models are not slowed down by solvent viscosity. For that purpose, we monitored how the end-to-end distance distributions of PEG6 starting from the all trans stretched conformer approach the limit distribution of the ES model  $g_{ES}^{(1\mu s)}(x)$  obtained with the maximum available time span of 1  $\mu$ s. The time evolutions of the integral square deviations  $[\Delta g^{(t)}]^2$ , eq 2, from limit distribution  $g_{ES}^{(1\mu s)}(x)$  are displayed in Figure 8 for the ES and IS GBSW models in a double logarithmic plot where the time decay appears approximately linear corresponding to a  $t^{-\alpha}$  power law with  $\alpha = 2/3$ .

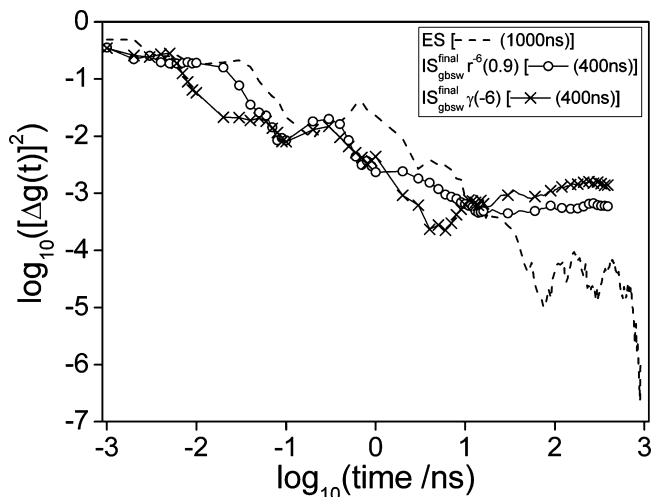
According to Figure 8, the end-to-end distance distribution of PEG6 seems to be converged after about 50 ns in the ES model (dashed line). At larger sampling times ( $t$ ),  $[\Delta g_{ES}^{(t)}]^2$  assumes a small plateau value with superimposed statistical fluctuations, since the ensemble of conformers is no longer large enough to exhibit further decay in  $[\Delta g_{ES}^{(t)}]^2$ . In the last 300 ns,  $[\Delta g_{ES}^{(t)}]^2$  decays to zero, since the limit distribution  $g_{ES}^{(1\mu s)}(x)$  representing the ensemble average is identical with the average using the maximum time span of 1  $\mu$ s.





**Figure 7.** End-to-end ( $C_1$ – $C_{19}$ ) distance distributions for PEG6 obtained from MD simulation data with ES (1  $\mu$ s) (dashed lines in parts A, B, C) and three IS models: (A) GBSW<sup>32–34</sup> (400 ns), (B) GBMV<sup>35,36</sup> (200 ns), and (C) FACTS<sup>37</sup> (400 ns) with varying surface energies, i.e., positive surface energy [ $\gamma = 30$  cal/(mol  $\text{\AA}^2$ )] (dotted lines), vanishing surface energy [ $\gamma = 0$  cal/(mol  $\text{\AA}^2$ )] (dashed-dotted lines). The solid lines in part A with  $\circ$  ( $\times$ ) symbols show the data from two optimized GBSW models (see Table 1).

The PEG6 conformer distributions that are based on the two IS GBSW models converge earlier at about 10 ns (solid lines in Figure 8). Beyond this time span until the maximum time span of 400 ns, the integral square deviations of the end-to-end distributions  $[\Delta g_{IS}^{(t)}]^2$  are approximately constant at a low value that corresponds to small deviations of the limit distribution  $g_{ES}^{(1\mu s)}(x)$  based on the ES model and the limit distributions  $g_{IS}^{(0.4\mu s)}(x)$  based on the two optimized IS GBSW models. Hence, the dynamics of PEG6 with the IS GBSW model are about a factor of 5 faster than with the corresponding ES model, in agreement with MD simulations



**Figure 8.** Efficiency of MD simulations to explore the conformational space of PEG6. Integral square deviation  $[\Delta g^{(t)}]^2$ , eq 2, of end-to-end atom pair distance distribution as a function of the time span used to obtain the ensemble average. Each time span starts at  $t = 0$ , where the initial PEG6 conformer is stretched all-trans. The lengths of the time spans to evaluate the averages of the distance distributions range from 1 ps to up to 1  $\mu$ s. Dashed line displays  $[\Delta g_{ES}^{(t)}]^2$  for the ES model. Solid lines show the time evolution of IS GBSW models with a maximum time span of 400 ns. The symbols  $\circ$  and  $\times$  refer to the two optimized IS GBSW models (see Table 1). The reference distribution  $g_{ES}^{(1\mu s)}(x)$  for all three cases considered here is taken from the ES model using the maximum time span of 1  $\mu$ s. Additional trajectories of 50 ns time were generated for PEG6 with different seeds of the random number generator for both optimized IS GBSW models, which show similar behavior (Figure S13 of Supporting Information).

of different molecules using the IS GBMV model with a low friction coefficient.<sup>27</sup> The exploration of the PEG6 conformational space as probed by the end-to-end distance distributions is surprisingly slow compared to the behavior of the autocorrelation function of the same quantity, which decays by 2 orders of magnitude faster in about 1 ns and 300 ps for the ES and IS GBSW models, respectively (see Figure S12 of the Supporting Information).

Snapshots of the end-to-end distance distributions for different time spans are shown in Figure S14 (Supporting Information) for the ES model of PEG6 together with the final distribution for the time span of 1  $\mu$ s. The end-to-end distance distribution of PEG6 obtained from MD simulation with the ES model where the 1  $\mu$ s trajectory was split into 10 segments, each 0.1  $\mu$ s long, is shown in Figure S15 of the Supporting Information.

## Conclusions

Explicit solvent (ES) MD simulations are computationally expensive due to the large number of atom pair interactions, which need to be evaluated. Alternatively, one can perform MD simulations with an implicit solvent (IS) model that compromises between efficiency and accuracy. In an IS model, the electrostatic solute–solvent interactions are often modeled by simplified GB electrostatics, and the hydrophobic



effect is modeled by a surface energy term. The latter should generally favor compact solute conformers.

In MD simulations with an ES model, the attractive vdW interactions between solute atoms compete with corresponding interactions of solute–solvent atom pairs, leading thus to balanced and effectively weaker solute–solute interactions. For small molecular units like DME, the conformers vary only moderately but markedly with surface energy. More compact conformers are obtained with increasing surface energy for the IS GBSW and GBMV models, while with the IS FACTS model, the opposite behavior was observed. Comparing 1  $\mu$ s MD simulation data of PEG with these three IS models, we found that the PEG conformers are much too compact for IS GBSW and GBMV models with positive surface energies. This phenomenon can be traced back to the lack of balance of attractive vdW interactions using these IS models with positive surface energy.

On the other hand, using these three GB models with vanishing surface energy yields PEG conformer distributions, which are globally realistic but may have some local deficiencies, which are similar as observed for DME. Adjusting the IS GBSW model to repair these deficiencies, it turned out that, in order to balance for a lack of solute–solvent attractive vdW interactions, the surface energy term must vanish and the attractive part of the vdW interactions must be reduced to 90% or alternatively, if the vdW interactions are not reduced, to use even a slightly negative surface energy. Hence, the influence from the unbalanced attractive vdW interactions is for PEG and DME even stronger than the hydrophobic effect. In addition, small but nevertheless significant reductions of the 1–4 and 1–5 O–H Coulomb interactions and adjustments of the torsion potentials were applied to obtain faithful local DME conformers. The former force field adjustments were necessary, since the IS GBSW model may have a tendency to overestimate Coulomb interactions at short distances. Interestingly, the force field originally adjusted for DME could be transferred to PEG unchanged.

The faithfulness of the PEG force field for an IS GBSW model developed in this contribution allows the study of entropy variations of PEG chain molecules under different geometry constraints (as for instance PEG bound with one or both ends on a wall or PEG moving through capillaries) most efficiently. For PEG6, the CPU time per time step is about a factor of 20 larger with the ES than with the IS GBSW model (see the discussion in the Supporting Information). MD simulations explore the conformational space of PEG6 in explicit and implicit water approximately according to a power law  $t^{-\alpha}$  in time  $t$  with exponent  $\alpha = 2/3$ . The dynamics of the end-to-end distance distribution of PEG6 obey a stretched exponential decay law for the ES model, while a power law was found for the IS GBSW model, demonstrating differences in the long term behavior between these models. Due to the lack of viscosity in the IS GBSW model, the PEG6 conformations are explored about five times faster with the IS GBSW model than with the ES model. Hence, in total, exploration of the conformational space of PEG6 is for the IS GBSW model about a factor of 100 faster than for the ES model.

**Acknowledgment.** We thank Dr. Martin Karplus for providing the program CHARMM, and we thank ZEDAT for providing generous access to ABACUS4, the high performance computing facility at the Freie Universitaet Berlin. We gratefully acknowledge helpful discussions with Gernot Kieseritzky. This work was supported by the Deutsche Forschungsgemeinschaft (Sfb 765 Project C1).

**Abbreviations.** DME, 1,2-dimethoxy-ethane; ES, explicit solvent; FACTS, fast analytical continuum treatment of solvation; GB, generalized Born; GBMV, GB using molecular volume; GBSW, GB with a simple switching; IS, implicit solvent; LJ interaction, Lennard-Jones interaction; MD, molecular dynamics; PEG, polyethylene glycol; SA, solvent accessible surface area; TGT/TGG'/TGG/TTT/, conformers with three consecutive dihedral angles (C–O–C–C, O–C–C–O, and C–C–O–C) centered around either 180° for trans (T) or +60° (G) or –60° (G') for gauche; vdW, van der Waals.

**Supporting Information Available:** Detailed description of explicit and implicit solvent simulation protocols; 2D structures of the 10 idealized DME and local PEG conformers; population of local DME conformers; End-to-end distance distribution of DME in IS<sub>gbsw</sub>, IS<sub>gbmv</sub>, IS<sub>facts</sub>; atom pair distance distributions of DME and PEG6; average free energies of torsion angles (C–O–C–C and O–C–C–O) of DME and PEG6; Distance distribution of DME and PEG6 with different IS models; average O–O distance distribution of PEG6; radius of gyration obtained from MD simulations of PEG6; autocorrelation function and integral square deviation, eq 2, of end-to-end distance of PEG6; end-to-end distance distribution of PEG6 in explicit solvent. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## References

- (1) Ben-Naim, A. Solvent Effects on Protein Stability and Protein Association. In *Protein-Solvent Interactions*, 1st ed.; Gregory, R. B., Ed.; Marcel Dekker, Inc.: New York, 1995; Vol. 592, pp 387–420.
- (2) *Water: A Comprehensive Treatise*; Franks, F., Ed.; Plenum Pub Corp: New York, 1972–1982; Vol. 1–7.
- (3) Eisenberg, D.; McLachlan, A. D. *Nature* **1986**, *319*, 199–203.
- (4) Teeter, M. M. *Annu. Rev. Biophys. Biophys. Chem.* **1991**, *20*, 577–600.
- (5) Soares, C. M.; Teixeira, V. H.; Baptista, A. M. *Biophys. J.* **2003**, *84*, 1628–1641.
- (6) Privalov, P. L.; Makhatadze, G. I. *J. Mol. Biol.* **1993**, *232*, 660–679.
- (7) Honig, B.; Yang, A.-S. Free Energy Balance in Protein Folding. In *Adv. Protein Chem.*; Anfinsen, C. B., Richards, F. M., Edsall, J. T., Eisenberg, D. S., Eds.; Academic Press: San Diego, 1995; Vol. 46, pp 27–58.
- (8) Klebe, G. *Drug Discovery Today* **2006**, *11*, 580–594.
- (9) Corbeil, C. R.; Moitessier, N. *J. Chem. Inf. Model.* **2009**, *49*, 997–1009.
- (10) Froloff, N.; Windemuth, A.; Honig, B. *Protein Sci.* **1997**, *6*, 1293–1301.

- (11) Frauenfelder, H.; Fenimore, P. W.; Chen, G.; McMahon, B. H. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 15469–15472.
- (12) Dill, K. A. *Biochemistry* **2002**, *29*, 7133–7155.
- (13) Ben-Naim, A. *Hydrophobic Interactions*; Plenum Press: New York: 1980.
- (14) Vaitheeswaran, S.; Yin, H.; Rasaiah, J. C.; Hummer, G. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 17002–17005.
- (15) Southall, N. T.; Dill, K. A.; Haymet, A. D. J. *J. Phys. Chem. B* **2001**, *106*, 521–533.
- (16) Makarov, V. A.; Feig, M.; Andrews, B. K.; Pettitt, B. M. *Biophys. J.* **1998**, *75*, 150–158.
- (17) Knapp, E. W.; Muegge, I. J. *J. Phys. Chem.* **1993**, *97*, 11339–11343.
- (18) Warwicker, J.; Watson, H. C. *J. Mol. Biol.* **1982**, *157*, 671–679.
- (19) Cramer, C. J.; Truhlar, D. G. *Chem. Rev.* **1999**, *99*, 2161–2200.
- (20) Roux, B.; Simonson, T. *Biophys. Chem.* **1999**, *78*, 1–20.
- (21) Gilson, M. K.; Davis, M. E.; Luty, B. A.; McCammon, J. A. *J. Phys. Chem.* **1993**, *97*, 3591–3600.
- (22) Baker, N. A. *Curr. Opin. Struct. Biol.* **2005**, *15*, 137–143.
- (23) Lopes, A.; Alexandrov, A.; Bathelt, C.; Archontis, G.; Simonson, T. *Proteins: Struct., Funct., Bioinf.* **2007**, *67*, 853–867.
- (24) Ferrara, P.; Apostolakis, J.; Caffisch, A. *Proteins: Struct., Funct., Bioinf.* **2002**, *46*, 24–33.
- (25) Chen, J.; Brooks, C. L. *Phys. Chem. Chem. Phys.* **2008**, *10*, 471–481.
- (26) Vorobjev, Y. N.; Almagro, J. C.; Hermans, J. *Proteins: Struct., Funct., Bioinf.* **1998**, *32*, 399–413.
- (27) Feig, M. *J. Chem. Theory Comput.* **2007**, *3*, 1734–1748.
- (28) Warshel, A.; Aqvist, J. *Annu. Rev. Biophys. Biophys. Chem.* **1991**, *20*, 267–298.
- (29) Kauzmann, W. Some Factors in the Interpretation of Protein Denaturation. In *Adv. Protein Chem.*; Anfinsen, C. B., Anson, M. L., Bailey, K., Edsall, J. T., Eds.; Academic Press: New York, 1959; Vol. 14, pp 1–63.
- (30) Tanford, C. Protein Denaturation. In *Adv. Protein Chem.*; Anfinsen, C. B., Anson, M. L., Edsall, J. T., Frederic, M. R., Eds.; Academic Press: New York, 1968; Vol. 23, pp 121–282.
- (31) Hummer, G.; Garde, S.; García, A. E.; Pohorille, A.; Pratt, L. R. *Proc. Natl. Acad. Sci. U. S. A.* **1996**, *93*, 8951–8955.
- (32) Im, W.; Lee, M. S.; Brooks, C. L. *J. Comput. Chem.* **2003**, *24*, 1691–1702.
- (33) Im, W.; Feig, M.; Brooks, C. L. *Biophys. J.* **2003**, *85*, 2900–2918.
- (34) Chen, J.; Im, W.; Brooks, C. L. *J. Am. Chem. Soc.* **2006**, *128*, 3728–3736.
- (35) Lee, M. S.; Salisbury, J. F. R.; Brooks, C. L., III *J. Chem. Phys.* **2002**, *116*, 10606–10614.
- (36) Lee, M. S.; Feig, M.; Salisbury, F. R.; Brooks, C. L. *J. Comput. Chem.* **2003**, *24*, 1348–1356.
- (37) Haberthür, U.; Caffisch, A. *J. Comput. Chem.* **2008**, *29*, 701–715.
- (38) Feig, M.; Onufriev, A.; Lee, M. S.; Im, W.; Case, D. A.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 265–284.
- (39) Im, W.; Beglov, D.; Roux, B. *Comput. Phys. Commun.* **1998**, *111*, 59–75.
- (40) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.
- (41) Lazaridis, T.; Karplus, M. *Proteins: Struct., Funct., Bioinf.* **1999**, *35*, 133–152.
- (42) Feig, M.; Brooks, C. L. *Curr. Opin. Struct. Biol.* **2004**, *14*, 217–224.
- (43) Chen, J.; Brooks, C. L.; Khandogin, J. *Curr. Opin. Struct. Biol.* **2008**, *18*, 140–148.
- (44) Tomasi, J. *Theor. Chem. Acc.* **2004**, *112*, 184–203.
- (45) Lee, B.; Richards, F. M. *J. Mol. Biol.* **1971**, *55*, 379–400.
- (46) Richmond, T. J. *J. Mol. Biol.* **1984**, *178*, 63–89.
- (47) Fraczekiewicz, R.; Braun, W. *J. Comput. Chem.* **1998**, *19*, 319–333.
- (48) Onufriev, A.; Case, D. A.; Bashford, D. *J. Comput. Chem.* **2002**, *23*, 1297–1304.
- (49) Zhu, J.; Shi, Y.; Liu, H. *J. Phys. Chem. B* **2002**, *106*, 4844–4853.
- (50) Mongan, J.; Simmerling, C.; McCammon, J. A.; Case, D. A.; Onufriev, A. *J. Chem. Theory Comput.* **2007**, *3*, 156–169.
- (51) Spassov, V. Z.; Yan, L.; Szalma, S. *J. Phys. Chem. B* **2002**, *106*, 8726–8738.
- (52) Wagoner, J. A.; Baker, N. A. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 8331–8336.
- (53) Gallicchio, E.; Kubo, M. M.; Levy, R. M. *J. Phys. Chem. B* **2000**, *104*, 6271–6285.
- (54) Floris, F.; Tomasi, J. *J. Comput. Chem.* **1989**, *10*, 616–627.
- (55) Pitera, J. W.; van Gunsteren, W. F. *J. Am. Chem. Soc.* **2001**, *123*, 3163–3164.
- (56) Gallicchio, E.; Levy, R. M. *J. Comput. Chem.* **2004**, *25*, 479–499.
- (57) Cheatham Iii, T. E.; Kollman, P. A. *Annu. Rev. Phys. Chem.* **2003**, *51*, 435–471.
- (58) Bailey, F. E.; Koleske, J. V. *Poly(ethylene oxide)*; Academic Press: New York, 1976.
- (59) *Poly(Ethylene Glycol) Chemistry: Biotechnical and Biomedical Applications*; Harris, J. M., Ed.; Plenum Publishing: New York, 1992.
- (60) Takahashi, Y.; Tadokoro, H. *Macromolecules* **1973**, *6*, 672–675.
- (61) Matsuura, H.; Miyazawa, T. *J. Polym. Sci. A-2* **1969**, *7*, 1735–1744.
- (62) Mueller-Plathe, F.; van Gunsteren, W. F. *Macromolecules* **1994**, *27*, 6040–6045.
- (63) Liu, H.; Müller-Plathe, F.; van Gunsteren, W. F. *J. Chem. Phys.* **1995**, *102*, 1722–1730.
- (64) Goutev, N.; Ohno, K.; Matsuura, H. *J. Phys. Chem. A* **2000**, *104*, 9226–9232.
- (65) Yoshida, H.; Kaneko, I.; Matsuura, H.; Ogawa, Y.; Tasumi, M. *Chem. Phys. Lett.* **1992**, *196*, 601–606.
- (66) Tasaki, K.; Abe, A. *Polym. J.* **1985**, *17*, 641–655.
- (67) Lusse, S.; Arnold, K. *Macromolecules* **1996**, *29*, 4251–4257.

- (68) Wang, R. L. C.; Kreuzer, H. J.; Grunze, M. *J. Phys. Chem. B* **1997**, *101*, 9767–9773.
- (69) Tasaki, K. *J. Am. Chem. Soc.* **1996**, *118*, 8459–8469.
- (70) Depner, M.; Schürmann, B. L.; Auriemma, F. *Mol. Phys.* **1991**, *74*, 715–733.
- (71) Heymann, B.; Grubmüller, H. *Chem. Phys. Lett.* **1999**, *307*, 425–432.
- (72) Pasut, G.; Guiotto, A.; Veronese, F. *Expert Opin. Ther. Pat.* **2004**, *14*, 859–894.
- (73) Mero, A.; Schiavon, O.; Pasut, G.; Veronese, F. M.; Emilitti, E.; Ferruti, P. *J. Bioact. Compat. Polym.* **2009**, *24*, 220–234.
- (74) Fuertges, F.; Abuchowski, A. *J. Controlled Release* **1990**, *11*, 139–148.
- (75) Harris, J. M.; Chess, R. B. *Nat. Rev. Drug Discovery* **2003**, *2*, 214–221.
- (76) Lascombe, M.-B.; Milat, M.-L.; Blein, J.-P.; Panabières, F.; Ponchet, M.; Prangé, T. *Acta Crystallogr., Sect. D* **2000**, *56*, 1498–1500.
- (77) Herrmann, A.; Arnold, K.; Pratsch, L. *Biosci. Rep.* **1985**, *5*, 689–696.
- (78) Zimmerberg, J.; Parsegian, V. A. *Nature* **1986**, *323*, 36–39.
- (79) O’Riordan, C. R.; Lachapelle, A.; Delgado, C.; Parkes, V.; Wadsworth, S. C.; Smith, A. E.; Francis, G. E. *Hum. Gene Ther.* **1999**, *10*, 1349–1358.
- (80) Prime, K. L.; Whitesides, G. M. *J. Am. Chem. Soc.* **1993**, *115*, 10714–10721.
- (81) Pertsin, A. J.; Grunze, M.; Garbuzova, I. A. *J. Phys. Chem. B* **1998**, *102*, 4918–4926.
- (82) Harder, P.; Grunze, M.; Dahint, R.; Whitesides, G. M.; Laibinis, P. E. *J. Phys. Chem. B* **1998**, *102*, 426–436.
- (83) Veronese, F.; Harris, J. M. *Adv. Drug Delivery Rev.* **2002**, *54*, 453–456.
- (84) Roberts, M. J.; Bentley, M. D.; Harris, J. M. *Adv. Drug Delivery Rev.* **2002**, *54*, 459–476.
- (85) Guillaudeau, S. J.; Fox, M. E.; Haidar, Y. M.; Dy, E. E.; Szoka, F. C.; Fréchet, J. M. J. *Bioconjugate Chem.* **2008**, *19*, 461–469.
- (86) Veronese, F.; Pasut, G. *Drug Discovery Today* **2005**, *10*, 1451–1458.
- (87) Mammen, M.; Choi, S.-K.; Whitesides, G. M. *Angew. Chem., Int. Ed.* **1998**, *37*, 2754–2794.
- (88) Kramer, R. H.; Karpen, J. W. *Nature* **1998**, *395*, 710–713.
- (89) Diestler, D. J.; Knapp, E. W. *Phys. Rev. Lett.* **2008**, *100*, 178101–4.
- (90) Diestler, D. J.; Knapp, E. W. *J. Phys. Chem. C* **2009**.
- (91) Oesterhelt, F.; Rief, M.; Gaub, H. E. *New J. Phys.* **1999**, *1*, 6.16.11.
- (92) Vorobyov, I.; Anisimov, V. M.; Greene, S.; Venable, R. M.; Moser, A.; Pastor, R. W.; Alexander, D.; MacKerell, J. *J. Chem. Theory Comput.* **2007**, *3*, 1120–1133.
- (93) Lee, H.; Venable, R. M.; MacKerell, A. D.; Pastor, R. W. *Biophys. J.* **2008**, *95*, 1590–1599.
- (94) Brooks, B. R.; Brooks, C. L.; MacKerell, A. D. J.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caffisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoseck, 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*, 1545–1614.
- (95) MacKerell, A. D. J.; Bashford, D.; Bellott, M. R. L.; Dunbrack, J.; 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.; W. E.; Reiher, I.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586–3616.
- (96) Neyertz, S.; Brown, D.; Thomas, J. O. *J. Chem. Phys.* **1994**, *101*, 10064–10073.
- (97) Smith, G. D.; Jaffe, R. L.; Yoon, D. Y. *J. Phys. Chem.* **1993**, *97*, 12752–12759.
- (98) Smith, G. D.; Jaffe, R. L.; Yoon, D. Y. *J. Am. Chem. Soc.* **1995**, *117*, 530–531.
- (99) Bedrov, D.; Borodin, O.; Smith, G. D. *J. Phys. Chem. B* **1998**, *102*, 5683–5690.
- (100) Masatoki, S.; Takamura, M.; Matsuura, H.; Kamogawa, K.; Kitagawa, T. *Chem. Lett.* **1995**, *24*, 991–992.
- (101) Begum, R.; Sagawa, T.; Masatoki, S.; Matsuura, H. *J. Mol. Struct.* **1998**, *442*, 243–250.
- (102) Born, M. *Z. Phys.* **1920**, *1*, 45–48.
- (103) Edinger, S. R.; Cortis, C.; Shenkin, P. S.; Friesner, R. A. *J. Phys. Chem. B* **1997**, *101*, 1190–1197.
- (104) Dominy, B. N.; Brooks, C. L. *J. Phys. Chem. B* **1999**, *103*, 3765–3773.
- (105) Bashford, D.; Case, D. A. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129–152.
- (106) Wesson, L.; Eisenberg, D. *Protein Sci.* **1992**, *1*, 227–235.
- (107) Weiner, S. J.; Kollman, P. A.; Nguyen, D. T.; Case, D. A. *J. Comput. Chem.* **1986**, *7*, 230–252.
- (108) Case, D. A.; Darden, T.; Gohlke, H.; Luo, R. K. M. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. *J. Comput. Chem.* **2005**, *26*, 1668–1688.
- (109) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (110) Hermann, R. B. *J. Phys. Chem.* **1972**, *76*, 2754–2759.
- (111) Durell, S. R.; Brooks, B. R.; Ben-Naim, A. *J. Phys. Chem.* **1994**, *98*, 2198–2202.
- (112) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.
- (113) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.
- (114) Nina, M.; Im, W.; Roux, B. *Biophys. Chem.* **1999**, *78*, 89–96.
- (115) Nosé, S.; Klein, M. L. *J. Chem. Phys.* **1983**, *78*, 6928–6939.
- (116) Hoover, W. G. *Phys. Rev. A* **1985**, *31*, 1695–1697.
- (117) Sitkoff, D.; Sharp, K. A.; Honig, B. *J. Phys. Chem.* **1994**, *98*, 1978–1988.
- (118) Knapp, E. W. *J. Comput. Chem.* **1992**, *13*, 793–798.

## Accurate Estimates of Free Energy Changes in Charge Mutations

Brittany R. Morgan<sup>†</sup> and Francesca Massi<sup>\*,‡</sup>

Department of Physics, Clark University, 950 Main Street, Worcester, Massachusetts 01610 and Department of Biochemistry and Molecular Pharmacology, University of Massachusetts, 55 Lake Avenue North, Worcester, Massachusetts 01655

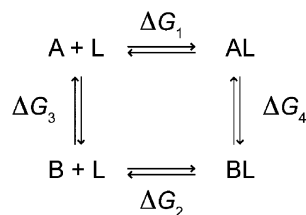
Received October 23, 2009

**Abstract:** The ability to determine the effect of charge changes on the free energy is necessary for fundamental studies of the electrostatic contribution to protein binding and stability. Currently, calculations of differences in free energy for charge mutations carried out in systems with periodic boundary conditions must include an approximate self-energy correction that can be on the same order of magnitude as the calculated free energy change. Here, a new method for accurately calculating the free energy change associated with any alchemical mutation, regardless of charge, is presented. In this method, paired mutations of opposite charge exactly cancel the self-energy term because of its quadratic charge dependence. Since the self-energy term implicitly cancels within the method, a correction never needs to be applied, and the statistical uncertainty associated is thereby removed. An implementation procedure is described and applied to the free energy of ionic hydration and a charged amino acid mutation.

### I. Introduction

The favorability of a chemical reaction is determined by the net change in free energy. Experimentally, the detailed mechanism by which these changes occur is not always apparent. The calculation of free energy differences using molecular simulations is a powerful tool to explore the atomistic basis of changes in free energy.<sup>1–3</sup> Free energy methods can be used to determine properties such as the free energy of solvation<sup>4–9</sup> or to explain and predict relative affinities in protein–ligand binding.<sup>10–16</sup> The physical principles that affect the free energy of binding are crucial for rational drug design and understanding the fundamental mechanisms of selectivity and affinity in binding. The importance of understanding the source of free energy differences emphasizes the need for accurate and widely applicable free energy calculation methods.

In an experiment, the free energies of binding for proteins A and B ( $\Delta G_1$  and  $\Delta G_2$  from Figure 1) are measured to determine the free energy difference ( $\Delta\Delta G$ ) of binding. The



**Figure 1.** Example of a thermodynamic cycle.  $\Delta G_1$  is the free energy of protein A binding a ligand (L), and  $\Delta G_2$  is the free energy of protein B binding L.  $\Delta G_3$  is the free energy of mutating protein A into protein B in the free state, and  $\Delta G_4$  is the free energy of mutating protein A into protein B in the bound state.

time required to compute  $\Delta\Delta G$  from spontaneous binding events is prohibitively long, however. Instead,  $\Delta\Delta G$  is calculated in free energy methods via the mutation of a protein A into B in the bound state and in the free state ( $\Delta G_3$  and  $\Delta G_4$  from Figure 1), where proteins A and B usually differ by a single amino acid.<sup>1–3,17–20</sup> This approach is possible due to the path independence of thermodynamic quantities, since

$$\Delta\Delta G = \Delta G_2 - \Delta G_1 = \Delta G_4 - \Delta G_3$$

\* Corresponding author phone: 508-856-4501; fax: 508-856-6464; e-mail: Francesca.Massi@umassmed.edu.

<sup>†</sup> Clark University.

<sup>‡</sup> University of Massachusetts.



Current free energy methods are limited by the need for potentially large corrections to the free energy change when mutations change the overall system charge; such corrections arise from the treatment of electrostatic potentials.<sup>21–26</sup> The correction can be on the same order of magnitude as the calculated free energy change, which makes these methods unattractive for non-neutral mutations due to concerns about accuracy.

The ability to determine the effect of charge changes on the free energy is necessary for fundamental studies of the electrostatic contribution to protein binding and stability. The method proposed here eliminates the need to explicitly include a correction, which removes the uncertainty in estimating the correction. This advancement gives free energy calculations a similar level of accuracy for any mutation, regardless of change in charge.

## II. Background

Free energy perturbation (FEP) and thermodynamic integration (TI)<sup>1–3,19,20</sup> are used in molecular dynamics (MD) or Monte Carlo (MC) simulations to determine free energy differences. The formalism for TI will be discussed, but the following arguments are equally applicable to FEP as well.

In TI, a scaling parameter  $\lambda$  is varied between 0 and 1, such that, given two states A and B, state A vanishes and state B grows as  $\lambda$  increases from 0 to 1. The free energy change of the mutation of state A into state B is calculated as

$$\Delta G = G_B - G_A \quad (1)$$

$$\Delta G = \int_0^1 \langle \Delta V \rangle_\lambda d\lambda \approx \sum_i \langle \Delta V \rangle_{\lambda_i} \Delta \lambda \quad (2)$$

where the energy difference  $\Delta V = V_B - V_A$  and  $\lambda$  satisfies  $V_\lambda = (1 - \lambda)V_A + \lambda V_B$ .<sup>19</sup> MD or MC is used to calculate the thermodynamic average  $\langle e^{-(H_B - H_A)} \rangle_A$  for each value of  $\lambda$ . The values of  $\lambda$  must be chosen carefully, to correctly evaluate  $\langle \Delta V \rangle_\lambda$ , as the accuracy of  $\Delta G$  depends on the accuracy of the thermodynamic average  $\langle \Delta V \rangle_\lambda$ .<sup>1–3</sup>

The potential energy calculated in an MD or MC simulation has both bonded and nonbonded contributions, but the electrostatic potential is particularly difficult to model both accurately and efficiently.<sup>21</sup> The long-range nature of the electrostatic potential makes boundary conditions and system size important factors in free energy calculations.<sup>27,28</sup> Over the past 20 years, calculations of the solvation free energy for atomic ions using finite or periodic boundary conditions have produced accurate results, independent of system size.<sup>4,28–35</sup> An important difference between the use of finite versus periodic boundary conditions lies in the fact that, while finite boundaries, such as finite droplets or semifinite slab systems, have a liquid–vacuum interface, infinite periodic boundary conditions (tin-foil) do not. For this reason, calculations of hydration free energies of atomic ions performed with finite boundary conditions include the contribution of the surface potential of this interface where bulk solvent simulations performed using periodic boundary conditions do not.<sup>25,27,36</sup> This interface potential is an

important contribution to the solvation free energy of charged solutes that arises when an ion crosses the liquid–vacuum boundary. Its contribution to the free energy of solvation is  $q\phi_{l-v}$ , where  $q$  is the charge of the solute and  $\phi_{l-v}$  is the electrostatic potential jump due to the liquid–vacuum interface, which depends upon the solvent model used in the simulation.<sup>37</sup> Free energies obtained in the absence of the interface potential are also called intrinsic free energies. The electrostatic interface potential is essential for obtaining accurate estimates of the free energy of hydration for ionic molecules, and its contribution must be added to the intrinsic free energy of hydration obtained from simulations that use tin-foil boundary conditions.<sup>25,36</sup> The magnitude of the interface potential depends on the particular electrostatic summation methodology employed, for example, whether the P-sum (particle-based) or M-sum (molecule-based) convention is used.<sup>25</sup>

For free energy calculations where the difference in the free energy change,  $\Delta\Delta G$ , is the desired quantity (Figure 1), the free energy change associated with crossing the liquid–vacuum interface is the same for  $\Delta G_3$  and  $\Delta G_4$ ; because the charge and the solvent model used are the same, the contributions that the interface potential makes will cancel in  $\Delta\Delta G$ . For this reason, intrinsic free energy changes, which do not include the interface potential, can be used to evaluate  $\Delta\Delta G$ . Calculations of  $\Delta\Delta G$  associated with the binding of two proteins, wild-type and mutant, to the same ligand (Figure 1) often use Ewald summation with tin-foil boundary conditions.<sup>38–40</sup>

The electrostatic potential in the Ewald summation is ill-defined if the system is not neutral. For charged solutes, the Ewald formulation implicitly neutralizes the charge with a homogeneous background charge.<sup>41</sup> This neutralizing plasma (NP) is a uniform charge distribution which is present everywhere in the system, including inside the solute. An alternative to the NP is to add explicit counterions to restore neutrality. Adding counterions is preferable because the NP is not an accurate physical representation of the charge density in solution.<sup>21</sup>

There is, however, an artifact introduced when using Ewald summation. Due to the periodicity of the Ewald implementation, charged atoms in the central cell have an electrostatic interaction with their periodic images, creating an artificial self-energy term which has a system size-dependent contribution to the free energy change.<sup>21,22</sup> In neutral systems, the contribution of the self-energy term to the free energy change is zero.

In TI or FEP, if states A and B have different charges, the self-energy term is different for each value of the scaling parameter,  $\lambda$ , during the mutation of state A into state B. For this reason, the self-energy term must then be accounted for in the final estimate of the free energy change. Hence, the intrinsic free energy change is equal to the calculated free energy change ( $\Delta G^{\text{calc}}$ ) plus a correction term,

$$\Delta G = \Delta G^{\text{calc}} + \text{correction} \quad (3)$$

The Ewald summation with tin-foil boundaries has no interface and is always neutral; thus, there is no contribution from the interface potential in the free energy change

calculated by TI or FEP.<sup>36</sup> The intrinsic free energy change calculated by TI or FEP using tin-foil boundary conditions can be used to obtain an estimate of  $\Delta\Delta G$  (Figure 1), noting that if the calculation involves a charged species, the interface potential has the same contribution to  $\Delta G_3$  and  $\Delta G_4$ . However, if we are interested in accurately determining the solvation free energy of a single ion, the contribution of the interface potential must be added to the intrinsic free energy change.<sup>25,36</sup> The interface potential is a physical quantity present in the experimental solvation free energies of charged solutes which must be added (in the case of Ewald summation with tin-foil boundaries) to the intrinsic free energy for accurate calculations of solvation free energies.

In contrast, the self-energy is an artifact of the infinite periodicity in the Ewald implementation, which is always present in non-neutral systems.<sup>21,22</sup> For accurate estimates of any intrinsic free energy change for charge mutations, the self-energy must be removed from the free energy change calculated by TI or FEP (eq 3). Self-energy artifacts in free energy calculations arising from mutations of charged solutes during TI or FEP have been well studied by calculating the free energy of hydration of atomic ions.<sup>23,24,31,32,42–45</sup>

A correction has been derived for the self-energy contribution to the free energy change when using Ewald summation with periodic boundary conditions.<sup>41</sup> The electrostatic potential ( $U_{\text{Coul}}$ ) of an infinitely periodic system is defined as

$$U_{\text{Coul}} = \sum_{1 \leq i \leq j \leq N} q_i q_j \phi_{\text{EW}}(\mathbf{r}_{ij}) + \frac{1}{2} \sum_{1 \leq i \leq N} q_i^2 \xi_{\text{EW}} \quad (4)$$

where  $r_{ij}$  is the distance between atoms  $i$  and  $j$  with charges  $q_i$  and  $q_j$ ,  $N$  is the total number of atoms, and  $\phi_{\text{EW}}(\mathbf{r})$  is the position-dependence of the electrostatic potential.  $\xi_{\text{EW}}$  is the self-energy term defined by the lattice summation used in Ewald summation,<sup>46–48</sup> which has the form

$$\xi_{\text{EW}} = \lim_{r \rightarrow 0} \left( \phi_{\text{EW}}(\mathbf{r}) - \frac{1}{r} \right) \quad (5)$$

$\xi_{\text{EW}}$  is the electrostatic potential from the periodic images and the neutralizing background charge in a Wigner lattice.<sup>46–48</sup> From eq 4, the energy difference between an initial state with charge  $q_0$  and a final state with charge  $q_1$  and can be calculated as

$$\Delta U_{\text{Coul}} = \Delta q \phi_{\text{EW}}(\mathbf{r}) + \frac{1}{2} \xi_{\text{EW}} (q_1^2 - q_0^2), \quad \text{where } \xi_{\text{EW}} = \frac{-2.837}{4\pi\epsilon_0 L} \quad (6)$$

for Ewald summations in a cubic lattice with side length  $L$  and  $\Delta q = q_1 - q_0$ .<sup>47,48</sup>

### III. Previous Work

The second term of eq 6 is the self-energy correction ( $C_{\text{EW}}$ ) which is added to the intrinsic free energy change calculated via TI or FEP ( $\Delta G^{\text{calc}}$ ) in eq 3, such that

$$\Delta G = \Delta G^{\text{calc}} + C_{\text{EW}} \quad (7)$$

Ideally,  $C_{\text{EW}}$  is equal to and opposite of the true self-energy ( $E_{\text{self}}$ ), but this approximation may not be accurate. Charge mutations are problematic even for single atom mutations,<sup>25,26</sup> and there are far fewer examples of charge mutations<sup>10–12,49</sup> than neutral mutations in biomolecular simulations. While charge mutations of atomic ions have been well studied, little has been done to advance the application to more complex systems.

The approach often used for charge mutations is to avoid the self-energy problem by performing two simultaneous mutations which render the system neutral in both initial and final states.<sup>10,12</sup> For neutral systems, the self-energy is zero, and no correction is needed ( $\Delta G = \Delta G^{\text{calc}}$ ). Either a residue far from the area of interest or a neutral dummy atom in the system can be mutated to compensate for the change in charge. Using this approach, the free energy change of the dual mutation is calculated ( $\Delta G_1 = \Delta G_1^{\text{calc}}$ ); hence the free energy change of the counter mutation must be known in order to obtain the free energy change of the charge mutation. Assuming the mutations are independent,  $\Delta G_1$  can be written as

$$\Delta G_1 = \Delta G_A + \Delta G_B \quad (8)$$

where  $\Delta G_A$  is the free energy change of the charge mutation and  $\Delta G_B$  is the free energy change of the counter mutation;  $\Delta G_A$  is the quantity of interest.

To find  $\Delta G_A$ , we must calculate  $\Delta G_B$  separately, but the counter mutation is also a charge mutation, and the same concerns regarding the self-energy apply in determining its free energy change. However, the self-energy is dependent on system size as well as charge (eq 6). If the system size is sufficiently large such that the self-interactions are minimal, then the self-energy vanishes and  $\Delta G_B = \Delta G_B^{\text{calc}}$ , but at what point the system size is sufficiently large can be difficult to quantify a priori.

To determine whether the self-interactions are minimal, the neutral dummy atom can be mutated into an atomic ion for the counter mutation (mutation B), and  $\Delta G_B$  can be compared to experimental free energies of hydration. As these free energy changes can be on the order of 100 kcal/mol, accuracy is an important concern. In order to reproduce experimental values, self-energy corrections to  $\Delta G_B$  dependent on the ionic charge, ionic radius, boundary conditions, system size, electrostatic scheme, and solvent model are needed,<sup>25,26</sup> and the liquid–vacuum interface potential must also be included.<sup>25,26,36</sup> The solvation free energy can be accurately calculated for atomic ions when these corrections are included.<sup>21–26</sup> However, the free energy of ionic hydration cannot be measured directly, and extrathermodynamic assumptions<sup>37</sup> introduce inconsistencies in the experimental values,<sup>50–56</sup> which makes it difficult to assess the accuracy of  $\Delta G_B$ .

Another serious concern exists with no current solution. For the self-energy to vanish, the charge mutation and the counter mutation must be performed simultaneously to preserve electrostatic neutrality. It is assumed that the free energy changes of the mutations are independent (eq 8). If they are not independent, then the free energy change of the dual mutation becomes

$$\Delta G_1^* = \Delta G_A + \Delta G_B + \Delta G_{AB} \quad (9)$$

where  $\Delta G_{AB}$  is the free energy change due to the interaction, and it is impossible to find  $\Delta G_A$  even with an accurate estimate of  $\Delta G_B$ .

The method of dual mutations is not widely used because of concerns regarding these assumptions. In this work, we present a method which also uses dual mutations to remove the self-energy but does not suffer from these limitations. In section IV, we present the theoretical rationale and three cases for which the self-energy correction is zero. These cases can be used to accurately find the free energy change of both the charge mutation and the counter mutation and, additionally, confirm the independence of the dual mutations. The assumption of a sufficiently large system size is not needed; in fact, the method presented in this work can directly quantify at what point the system size becomes sufficiently large to make the self-energy negligible. More importantly, this method is directly applicable to complex systems, and independence can be verified in any system. It is not limited to the basic examples used to illustrate the method in section V.C. A detailed procedure for applying the method is presented in section V.A and demonstrated by two examples in section V.C.

#### IV. Theoretical Basis

The self-energy correction ( $C_{EW}$ ) from eq 6 can be rewritten as

$$C_{EW} = \frac{1}{2}\xi_{EW}(q_1^2 - q_0^2) = \frac{1}{2}\xi_{EW}(q_1 - q_0)(q_1 + q_0) \quad (10)$$

In this form, it becomes obvious that  $C_{EW} = 0$  not only for  $(q_0, q_1) = (0, 0)$ , but also for  $(a, -a)$  and  $(a, a)$ . Therefore, for any mutation where  $|q_1| = |q_0|$ , the self-energy correction  $C_{EW}$  vanishes. In the remainder of this section, we will discuss explicitly how  $C_{EW}$  depends on  $q_0$  and  $q_1$ .

**Case 1A.** Consider the case when  $(q_0, q_1) = (0, 0)$  as the mutation  $A^0 \rightarrow B^0$ , where initial state  $A^{q_0}$ , with charge  $q_0 = 0$ , mutates into final state  $B^{q_1}$ , with charge  $q_1 = 0$ . Substitution of  $q_0$  and  $q_1$  into eq 10 yields the expected result of

$$C_{EW} = \frac{1}{2}\xi_{EW}(q_1^2 - q_0^2) = \frac{1}{2}\xi_{EW}((0)^2 - (0)^2) = 0 \quad (11)$$

**Case 1B.** The dual mutation  $A^0 + B^0 \rightarrow A^+ + B^-$  also satisfies  $(q_0, q_1) = (0, 0)$ ; hence,  $C_{EW} = 0$ . If these mutations are separated into two single mutations, (1)  $A^0 \rightarrow A^+$  and (2)  $B^0 \rightarrow B^-$ , with  $C_{EW} = C_{EW}^{(1)} + C_{EW}^{(2)}$ , then

$$C_{EW}^{(1)} = \frac{1}{2}\xi_{EW}((1)^2 - (0)^2) = \frac{1}{2}\xi_{EW} \quad (12)$$

and

$$C_{EW}^{(2)} = \frac{1}{2}\xi_{EW}((-1)^2 - (0)^2) = \frac{1}{2}\xi_{EW} \quad (13)$$

$C_{EW}$  is no longer equal to zero if the two mutations are not performed simultaneously.

**Case 2A.** As an example of the case  $(q_0, q_1) = (a, -a)$ , consider the mutation  $A^+ \rightarrow B^-$ . Substitution of  $q_0 = +1$  and  $q_1 = -1$  into eq 10 yields

$$C_{EW} = \frac{1}{2}\xi_{EW}(q_1^2 - q_0^2) = \frac{1}{2}\xi_{EW}((-1)^2 - (1)^2) = \frac{1}{2}\xi_{EW}(1 - 1) = 0 \quad (14)$$

Thus, the Ewald implementation considers the self-energy term of the mutation  $A^+ \rightarrow B^-$  equivalent to  $A^0 \rightarrow B^0$ .

Thermodynamic cycles are path-independent, which makes it possible to decompose the mutation  $A^+ \rightarrow B^-$  into two serial mutations, (1)  $A^+ \rightarrow N^0$  and (2)  $N^0 \rightarrow B^-$ , where  $N^0$  is some neutral species. For the first mutation

$$C_{EW}^{(1)} = \frac{1}{2}\xi_{EW}((0)^2 - (1)^2) = -\frac{1}{2}\xi_{EW} \quad (15)$$

and for the second mutation

$$C_{EW}^{(2)} = \frac{1}{2}\xi_{EW}((-1)^2 - (0)^2) = \frac{1}{2}\xi_{EW} \quad (16)$$

$C_{EW}^{(1)}$  is equal and opposite to  $C_{EW}^{(2)}$ , demonstrating that  $C_{EW} = C_{EW}^{(1)} + C_{EW}^{(2)} = 0$  as well after decomposition.

**Case 2B.** The dual mutation  $A^+ + B^0 \rightarrow A^0 + B^-$  has the same charge at the initial and final states as the mutation  $A^+ \rightarrow B^-$ , which was shown to have  $C_{EW} = 0$  in eq 14. The mutation  $A^+ + B^0 \rightarrow A^0 + B^-$  can be separated into two single mutations (1)  $A^+ \rightarrow A^0$  and (2)  $B^0 \rightarrow B^-$ , as in case 1B, but here  $C_{EW} = C_{EW}^{(1)} + C_{EW}^{(2)} = 0$  as in eqs 15 and 16.

**Case 3A.** The mutation  $A^+ \rightarrow B^+$  corresponds to the case  $(q_0, q_1) = (a, a)$ . Substituting into eq 7,

$$C_{EW} = \frac{1}{2}\xi_{EW}(q_1^2 - q_0^2) = \frac{1}{2}\xi_{EW}((1)^2 - (1)^2) = 0 \quad (17)$$

**Case 3B.** Separation of  $A^+ + B^0 \rightarrow A^0 + B^+$  into (1)  $A^+ \rightarrow A^0$  and (2)  $B^0 \rightarrow B^+$  results in the same cancellation of  $C_{EW}^{(1)}$  and  $C_{EW}^{(2)}$  as in case 2B.

Cancellation of  $C_{EW}$  occurs for cases 2B and 3B, but not 1B, which leads to the conclusion that the *direction* of the change in charge is the basis of the cancellation, while the sign of the individual charges is irrelevant. This property is akin to the ionic strength, as both depend on the magnitude of the charge in the system, not the sign.

The cancellation of  $C_{EW}$  results from the quadratic dependence of  $C_{EW}$ , which arises because the Coulomb potential is quadratic in the charge. Any energy derived from a point-charge Coulomb potential will be quadratic in the charge, including the self-energy; hence, the self-energy will also cancel according to the preceding argument, provided that other dependencies in the self-energy have a negligible effect compared to the charge dependence.<sup>25</sup> This cancellation does not occur, however, for the interface potential (discussed in section II), which is linear in the charge.

The problems inherent to charge mutations are well understood for atomic ions.<sup>21-26,36,57</sup> Unfortunately, the translation of these findings to more complex systems is not straightforward. The work on atomic ions shows dependencies in the free energy change on ionic charge, ionic radius, boundary conditions, system size, electrostatic scheme, and the solvent model.<sup>25</sup> The method presented here addresses the charge dependence, but self-energy corrections for infinitely periodic systems still include a dependence on the

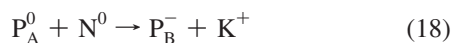


solute radius.<sup>21,26,58</sup> We will demonstrate in section V.C that the dependence on solute radius is weak and does not reduce the applicability of this method. The effect of system size will also be discussed.

## V. Application to Charge Mutations in Proteins

The cancellation demonstrated for  $C_{EW}$  in section IV, which applies to the self-energy ( $E_{self}$ ) as well, can be used to accurately find the free energy change of both the charge mutation and the counter mutation and, additionally, confirm the independence of the mutations. We will use FEP to outline and demonstrate the method, using the notation  $\Delta G^{FEP} = \Delta G^{calc}$ .

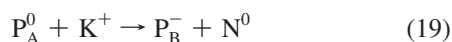
Consider the change in free energy for protein  $P_A$  which begins neutral and becomes protein  $P_B$  with a charge of  $-1$  through the mutation of a residue ( $P_A^0 \rightarrow P_B^-$ ). Concurrently performing the mutation  $N^0 \rightarrow K^+$ , where  $N^0$  is a neutral dummy atom and  $K^+$  is a potassium ion, renders the system charge neutral continuously from the initial to the final state, as in section III. This is described by the dual mutation



with free energy change  $\Delta G_1$ .

As shown in cases 1A and 1B,  $C_{EW} = 0$  only when the two mutations are performed simultaneously, and thus  $\Delta G_1 = \Delta G_1^{FEP}$  (eq 7). Assuming that the two concurrent mutations are independent, the calculated free energy change of the dual mutation in eq 18 ( $\Delta G_1^{FEP}$ ) is the combined free energy change of the two single mutations without a contribution from  $E_{self}$  such that  $\Delta G_1^{FEP} = \Delta G_{P_A^0 \rightarrow P_B^-} + \Delta G_{N^0 \rightarrow K^+}$  (eq 8), where  $\Delta G_{P_A^0 \rightarrow P_B^-}$  is the free energy change of the mutation  $P_A^0 \rightarrow P_B^-$  and  $\Delta G_{N^0 \rightarrow K^+}$  is the free energy change of the mutation  $N^0 \rightarrow K^+$ . From this calculation, we only obtain the sum of  $\Delta G_{P_A^0 \rightarrow P_B^-}$  and  $\Delta G_{N^0 \rightarrow K^+}$ .

However, we can avoid the problems outlined in section III by additionally performing the mutation



with free energy change  $\Delta G_2$ , which results in  $C_{EW} = 0$  whether performed concurrently in the same FEP or separately as shown in cases 2A and 2B. The direction of the mutation of the potassium ion is reversed ( $K^+ \rightarrow N^0$ ), which has the opposite free energy change ( $\Delta G_{K^+ \rightarrow N^0}^{FEP} = -\Delta G_{N^0 \rightarrow K^+}^{FEP}$ ). The advantage is that we can calculate  $\Delta G_2^{FEP}$  of the dual mutation in eq 19, which is also free of  $E_{self}$  (i.e.,  $\Delta G_2 = \Delta G_2^{FEP}$ ), but in this case,  $\Delta G_2^{FEP} = \Delta G_{P_A^0 \rightarrow P_B^-} - \Delta G_{N^0 \rightarrow K^+}$ . Combining  $\Delta G_1^{FEP}$  and  $\Delta G_2^{FEP}$ , we find the individual free energy changes  $\Delta G_{P_A^0 \rightarrow P_B^-}$  and  $\Delta G_{N^0 \rightarrow K^+}$  as

$$\Delta G_{P_A^0 \rightarrow P_B^-} = \frac{\Delta G_1^{FEP} + \Delta G_2^{FEP}}{2} \text{ and} \\ \Delta G_{N^0 \rightarrow K^+} = \frac{\Delta G_1^{FEP} - \Delta G_2^{FEP}}{2} \quad (20)$$

As already discussed in case 2B, the second mutation (eq 16) can also be performed as two separate mutations, (1)  $P_A^0$

$\rightarrow P_B^-$  and (2)  $K^+ \rightarrow N^0$ , to give  $\Delta G_{P_A^0 \rightarrow P_B^-}^{FEP}$  and  $-\Delta G_{N^0 \rightarrow K^+}^{FEP}$ ; we can add  $\Delta G_{P_A^0 \rightarrow P_B^-}^{FEP}$  and  $-\Delta G_{N^0 \rightarrow K^+}^{FEP}$  to obtain  $\Delta G_2$ . If the two mutations are independent, this  $\Delta G_2$  should be equal to  $\Delta G_2^{FEP}$  calculated from the FEP of eq 19.

By the method presented here, we can obtain  $\Delta G_{P_A^0 \rightarrow P_B^-}$  without using a self-energy correction or needing to calculate  $\Delta G_{N^0 \rightarrow K^+}$  separately, as has previously been necessary. The self-energy is a simulation artifact that needs to be removed. We do this by performing the free energy calculation on a modified system that includes an additional ion, as illustrated in eqs 18 and 19, to cancel the self-energy. The undesired quantity ( $\Delta G_{N^0 \rightarrow K^+}$ ) can be determined and removed from the calculated  $\Delta G$  (as shown in eq 20), whereas the self-energy term associated with the calculated free energy for the mutation of the charged species of interest alone could not. In this method, only self-consistency is necessary for the atomic ion mutations; i.e., force field, solvent model, electrostatic convention, and boundary conditions should be identical.

The method uses the equality  $\Delta G_{K^+ \rightarrow N^0}^{FEP} = -\Delta G_{N^0 \rightarrow K^+}^{FEP}$ , which is true if the hysteresis of the forward and reverse FEP is negligible;<sup>59</sup> this is a reasonable assumption for atomic ion mutations, as used here, and can be easily verified. Moreover, the crucial assumption of independence of the two mutations can now be confirmed, as detailed below.

In order to guarantee the independence of the two concurrent mutations, it is necessary to position the counterion at an adequate distance from the protein in order to minimize the electrostatic interaction between the protein and the ion. The Bjerrum length is defined as the distance at which the Coulomb interaction between two monovalent ions in a uniform dielectric is equal to the thermal energy,  $k_B T$ .<sup>60</sup> In water at 300 K, the Bjerrum length is  $\sim 7$  Å. This distance can be used to approximately determine how far apart the protein and the counterion should be placed in the simulation. For example, for free energy calculations of proteins in water at 300 K, a distance of  $\sim 7$  Å or greater between the surface of the protein and the counterion will be adequate to ensure that the electrostatic interaction between the protein and the ion is effectively screened.

**A. Procedure to Calculate the Free Energy Change of Any Charge Mutation.** In the example above, the FEP of the protein mutation was performed three times: as the mutations  $P_A^0 + N^0 \rightarrow P_B^- + K^+$ ,  $P_A^0 + K^+ \rightarrow P_B^- + N^0$ , and  $P_A^0 \rightarrow P_B^-$ . To reduce the computational demand of this method, we propose the following generalized procedure to calculate the free energy change for any charge mutation.

**Step 1:** Determine  $\Delta G$  of a counter mutation such as an atomic ion ( $K^+$ , in this example).

**Step 2:** Combine  $\Delta G_{N^0 \rightarrow K^+}$  with  $\Delta G_2^{FEP} = \Delta G_{P_A^0 \rightarrow P_B^-} - \Delta G_{N^0 \rightarrow K^+}$  of the dual mutation  $P_A^0 + K^+ \rightarrow P_B^- + N^0$  to find  $\Delta G_{P_A^0 \rightarrow P_B^-}$ .

**Step 3:** Use the single mutations  $P_A^0 \rightarrow P_B^-$  and  $K^+ \rightarrow N^0$  to confirm the independence of the mutations.

Although only the first two steps are necessary to find the free energy change, we perform all three steps to check the accuracy of the method and the independence of the two mutations.



**Table 1.** Free Energy Change of the Mutations  $N^0 \rightarrow Cl^-$  and  $N^0 \rightarrow K^+$ 

step <sup>a</sup>	$\Delta G^{FEP}$	mutation	
1-1: $\Delta G_1^{FEP} = \Delta G_{N^0 \rightarrow Cl^-} + \Delta G_{N^0 \rightarrow K^+}$	$-162.8 \pm 0.7$	$N^0 + N^0 \rightarrow Cl^- + K^+$	
1-2: $\Delta G_2^{FEP} = \Delta G_{N^0 \rightarrow Cl^-} - \Delta G_{N^0 \rightarrow K^+}$	$-21.5 \pm 0.6$	$N^0 + K^+ \rightarrow Cl^- + N^0$	
3-3: $\Delta G_{N^0 \rightarrow Cl^-}^{FEP} - \Delta G_{N^0 \rightarrow K^+}^{FEP}$	$-21.3 \pm 0.4$	$N^0 \rightarrow Cl^-$ $K^+ \rightarrow N^0$	
$\Delta G_{N^0 \rightarrow Cl^-}^{FEP}$ <sup>b</sup>	$-91.9 \pm 0.3$	$\Delta G_{N^0 \rightarrow K^+}^{FEP}$ <sup>b</sup>	$-70.6 \pm 0.1$
$\Delta G_{N^0 \rightarrow Cl^-}$ <sup>c</sup>	$-92.2 \pm 1.3$	$\Delta G_{N^0 \rightarrow K^+}$ <sup>c</sup>	$-70.7 \pm 1.3$

<sup>a</sup> Mean and error of free energy change from FEP calculations ( $\Delta G^{FEP}$ ) in kcal/mol. The mutations used to obtain each  $\Delta G^{FEP}$  are listed, see section V for details. The values of the independent mutations used to calculate the free energy change for step 3-3 are reported in the next line (b). <sup>b</sup> Mean and error of the hydration free energy computed directly using FEP on the individual ions. <sup>c</sup> Mean and error of final free energy change ( $\Delta G$ ) in kcal/mol for the mutation of a neutral dummy atom ( $N^0$ ) into chloride ( $Cl^-$ ) and potassium ( $K^+$ ) ions calculated from steps 1-1 and 1-2 according to eq 20.

*Step 1: Calculate the Free Energy Change of Counter Mutations.* The free energy change of the counter mutation is a constant value and only needs to be determined once for a given water model.<sup>61</sup> Atomic ions are convenient for the counter mutation because the FEP can be performed in a short time compared to the mutation of a residue in a solvated protein. The system size used in the simulations should be sufficiently large to minimize interactions between the mutating ions, which is verified in step 3.

As an example, we will calculate the free energy change of chloride and potassium ions associated with the mutations  $N^0 \rightarrow Cl^-$  ( $\Delta G_{N^0 \rightarrow Cl^-}$ ) and  $N^0 \rightarrow K^+$  ( $\Delta G_{N^0 \rightarrow K^+}$ ), using the following three step strategy. In each step,  $E_{self}$  is identically zero. The numerical results are presented in Table 1.

1-1. Calculate  $\Delta G_1^{FEP} = \Delta G_{N^0 \rightarrow Cl^-} + \Delta G_{N^0 \rightarrow K^+}$  of the mutation  $N^0 + N^0 \rightarrow Cl^- + K^+$ .

1-2. Calculate  $\Delta G_2^{FEP} = \Delta G_{N^0 \rightarrow Cl^-} - \Delta G_{N^0 \rightarrow K^+}$  of the mutation  $N^0 + K^+ \rightarrow Cl^- + N^0$ .

1-3. The free energy changes  $\Delta G_{N^0 \rightarrow Cl^-}$  and  $\Delta G_{N^0 \rightarrow K^+}$  can be determined from  $\Delta G_1^{FEP}$  and  $\Delta G_2^{FEP}$  as in eq 20.

*Step 2: Calculate the Free Energy Change of Any Mutation.* The free energy change of the mutation  $P_A^0 \rightarrow P_B^-$  ( $\Delta G_{P_A^0 \rightarrow P_B^-}$ ) can now be calculated using  $\Delta G_{N^0 \rightarrow K^+}$  from step 1-3. The mutation  $P_A^0 + K^+ \rightarrow P_B^- + N^0$  is used here because, while it would be equally valid mathematically to use the mutation  $P_A^0 + N^0 \rightarrow P_B^- + K^+$ , the first mutation reduces the interactions between the two charged species.

2-1. Calculate  $\Delta G_2^{FEP} = \Delta G_{P_A^0 \rightarrow P_B^-} - \Delta G_{N^0 \rightarrow K^+}$  of the dual mutation  $P_A^0 + K^+ \rightarrow P_B^- + N^0$ .

2-2. Use  $\Delta G_2^{FEP}$  and  $\Delta G_{N^0 \rightarrow K^+}$  (from step 1-3) to obtain  $\Delta G_{P_A^0 \rightarrow P_B^-}$  using the relationship  $\Delta G_{P_A^0 \rightarrow P_B^-} = \Delta G_2^{FEP} + \Delta G_{N^0 \rightarrow K^+}$ .

To find the free energy change for a positive protein mutation, use  $Cl^-$  instead of  $K^+$ . No special techniques are needed to calculate the free energy change of a continuous  $P_A^- \rightarrow P_B^+$  mutation, since the self-energy exactly cancels during the mutation (as in case 2A).

*Step 3: Confirm Independence of Concurrent Mutations.* We need to verify that, for the dual mutation  $P_A^0 + K^+ \rightarrow P_B^- + N^0$ ,  $\Delta G_2^{FEP} = \Delta G_{P_A^0 \rightarrow P_B^-} - \Delta G_{N^0 \rightarrow K^+}$  with no additional contribution from interactions between the two mutations. To do this, we will use the FEP of (1)  $P_A^0 \rightarrow P_B^-$  and (2)  $K^+ \rightarrow N^0$  in two separate simulations of the same system size ( $\xi_{EW}$  depends on  $L$ , eq 6). Steps 3-1 and 3-2 only need to be performed once per mutation because they are independent.

3-1. Calculate  $\Delta G_{P_A^0 \rightarrow P_B^-}^{FEP} = \Delta G_{P_A^0 \rightarrow P_B^-} + E_{self}$  of the single mutation  $P_A^0 \rightarrow P_B^-$ .

3-2. Calculate  $\Delta G_{N^0 \rightarrow K^+}^{FEP} = \Delta G_{N^0 \rightarrow K^+} + E_{self}$  of the single mutation  $N^0 \rightarrow K^+$ .

3-3. If  $\Delta G_{P_A^0 \rightarrow P_B^-}^{FEP} - \Delta G_{N^0 \rightarrow K^+}^{FEP} = \Delta G_2^{FEP}$  (from step 2-1), then the self-energy terms correctly cancel, and the two mutations are independent.

Furthermore, we can determine  $E_{self}$  using  $\Delta G_{N^0 \rightarrow K^+}$  from step 1-3, since  $E_{self} = \Delta G_{N^0 \rightarrow K^+}^{FEP} - \Delta G_{N^0 \rightarrow K^+}$  if the mutations used to determine  $\Delta G_{N^0 \rightarrow K^+}^{FEP}$  are independent. Knowledge of  $E_{self}$  can provide a metric to assess the amount of self-interactions in the simulation, as discussed below.

## B. Advantages of the Approach

*1. Increased Accuracy of the Free Energy Calculation.* The self-energy is implicitly canceled in the method presented here, which removes the need for a correction. The self-energy correction is no longer a crucial component in obtaining an accurate estimate of the free energy change for charge mutations.

*2. Knowledge of the Self-Energy.*  $E_{self}$  is a measure of the self-interactions between periodic images which depends on system size. For accurate MD simulations using PME, the system size should be sufficiently large so as to minimize these interactions, but what constitutes ‘‘sufficiently large’’ for long-range electrostatic interactions is not well-defined.  $E_{self}$  could be used as a quantitative metric to evaluate the system size.

*3. Adaptable to Changes of Larger Magnitude.* In the examples, the change in charge is limited to  $\pm 1$ . While it is possible to perform a mutation such as  $A^{2-} \rightarrow A^{2+}$  directly, as the extent of the change between the initial and final states increases, it becomes more difficult to obtain an accurate thermodynamic average for each value of  $\lambda$  (eq 2).<sup>1</sup> Decreasing the incremental change in  $\lambda$  can provide greater accuracy; however, the charge mutation can also be divided into smaller steps. A mutation  $A^{q-} \rightarrow A^{q+}$  can be decomposed into  $N$  steps, where  $2q/N$  can be less than 1 if fractional charge increments are used. As long as the incremental charge changes are identical in each direction, the self-energy still vanishes in the final free energy. Any staging method that can be applied to  $\lambda$  to increase the accuracy of the calculation<sup>62</sup> can be applied to  $q$  as well, but as  $q$  is not necessarily linear in  $\lambda$ , it may be easier to apply an efficient staging scheme as a function of  $q$  rather than  $\lambda$ .

*4. Identity of the Counter Mutation Is Irrelevant.* Since the contribution of the counter mutation is removed, the

**Table 2.** Free Energy Change of the Mutations  $D^- \rightarrow S^0$  and  $N^0 \rightarrow K^+$ 

step <sup>a</sup>	$\Delta G^{\text{FEP}}$	mutation
1-1: $\Delta G_{D^- \rightarrow S^0}^{\text{FEP}} = \Delta G_{D^- \rightarrow S^0}^{\text{FEP}} - \Delta G_{N^0 \rightarrow K^+}^{\text{FEP}}$	$82.5 \pm 1.3$	$D^- + N^0 \rightarrow S^0 + K^+$
1-2: $\Delta G_{D^- \rightarrow S^0}^{\text{FEP}} = \Delta G_{D^- \rightarrow S^0}^{\text{FEP}} + \Delta G_{N^0 \rightarrow K^+}^{\text{FEP}}$	$224.2 \pm 0.2$	$D^- + K^+ \rightarrow S^0 + N^0$
3-3: $\Delta G_{D^- \rightarrow S^0}^{\text{FEP}} - \Delta G_{N^0 \rightarrow K^+}^{\text{FEP}}$	$81.8 \pm 0.7$	$D^- \rightarrow S^0$ $N^0 \rightarrow K^+$
$\Delta G_{D^- \rightarrow S^0}^b$	$153.4 \pm 1.5$	$\Delta G_{N^0 \rightarrow K^+}^b$ $-70.9 \pm 1.5$

<sup>a</sup> Mean and error of free energy change from FEP calculations ( $\Delta G^{\text{FEP}}$ ) in kcal/mol. The mutations used to obtain each  $\Delta G^{\text{FEP}}$  are listed, see section V for details. <sup>b</sup> Mean and error of final free energy change ( $\Delta G$ ) in kcal/mol for the mutation of aspartic acid ( $D^-$ ) into serine ( $S^0$ ) and a neutral dummy atom ( $N^0$ ) into a potassium ion ( $K^+$ ).

**Table 3.** Self-Energy and Calculated Corrections Terms (kcal/mol)<sup>a</sup>

$L$ (Å)	$\Delta G^{\text{FEP}}$	$E_{\text{self}}$	$C_{\text{EW}}^b$	$\Delta \Delta G_{\text{solv}}^c$
		$K^+$		
29.16	$-70.6 \pm 0.1$	$0.1 \pm 1.4$	-16.2	-0.6
18.93	$-69.9 \pm 0.2$	$0.9 \pm 1.5$	-24.9	-1.6
11.13	$-65.4 \pm 0.2$	$5.3 \pm 1.5$	-42.3	-6.6
		$Cl^-$		
29.11	$-91.9 \pm 0.3$	$0.3 \pm 1.6$	-16.2	-0.8
18.97	$-90.8 \pm 0.1$	$1.4 \pm 1.4$	-24.8	-2.4
10.78	$-88.0 \pm 0.2$	$4.2 \pm 1.5$	-43.7	-11.2

<sup>a</sup> Mean and errors for the free energy calculated via FEP ( $\Delta G^{\text{FEP}}$ ) and the self energy ( $E_{\text{self}}$ ) for the mutation of a neutral dummy atom ( $N^0$ ) into a potassium ( $K^+$ ) or chloride ( $Cl^-$ ) ion for three system sizes ( $L$ ). The error of  $E_{\text{self}}$  is the error of  $\Delta G^{\text{FEP}}$  and  $\Delta G$  from Table 1. <sup>b</sup>  $C_{\text{EW}}$  is dependent on system size and charge (eq 10).<sup>41</sup> <sup>c</sup>  $\Delta \Delta G_{\text{solv}}$  is dependent on system size, charge, and solute radius (eq 21).<sup>21</sup>

accuracy of the estimated free energy change of the counter mutation does not affect the overall accuracy of the method. Moreover, unphysical mutations, such as fractional charges or artificial ions, can be used.

**C. Demonstration of the Method.** The method described above is illustrated here by several examples. FEP calculations were done with the molecular dynamics package NAMD 2.7<sup>63</sup> using the CHARMM 27<sup>64</sup> force field and explicitly solvated with the TIP3P water model. Each FEP was repeated five times to give the results in Tables 1–3. All simulations use PME with tin-foil boundary conditions. The free energy changes reported below are the intrinsic free energy change. The particular choice of boundary conditions affects the results, as discussed in section II. While the free energy change associated with the dual mutation reaction illustrated in eq 18 has no contribution from the surface potential term, the free energy change associated with the reaction illustrated in eq 19 does. If single ion solvation free energies are the goal of the calculation, then the contribution of the interface potential jump must be included.<sup>26,36,57</sup>

1. *Free Energy Change of Atomic Ion Mutations.* The free energy changes for  $N^0 \rightarrow Cl^-$  and  $N^0 \rightarrow K^+$  were calculated according to the procedure in step 1.  $N^0$  is a neutral dummy atom with the following Lennard-Jones parameters:  $\epsilon = 0$  and  $\sigma = 0$ . As a practical matter, before charging and alchemical mutation to the target species, the  $N^0$  state was transformed into an argon-like particle with the CHARMM 27<sup>64</sup> Lennard-Jones parameters for sodium (and zero charge). The free energy associated with insertion of this argon-like particle was  $2.0 \pm 0.2$  kcal/mol. For concurrent dual mutations, the two ions were fixed in the simulated water box at a distance  $L/2$  apart in  $x$ ,  $y$ , and  $z$  ( $L = 30$  Å). When

performing single mutations, the ion was fixed at the center of the simulation box ( $L = 30$  Å). The results from each step are listed in Table 1.

The free energy changes obtained from the concurrent dual mutations of steps 1-1 and 1-2 using eq 20 are  $\Delta G_{N^0 \rightarrow K^+} = -70.7 \pm 1.3$  kcal/mol and  $\Delta G_{N^0 \rightarrow Cl^-} = -92.2 \pm 1.3$  kcal/mol (Table 1, step 1-3).  $\Delta G_2^{\text{FEP}} = -21.5 \pm 0.6$  kcal/mol from the concurrent dual mutations (Table 1, step 1-2), which agrees well with  $\Delta G_{N^0 \rightarrow Cl^-}^{\text{FEP}} - \Delta G_{N^0 \rightarrow K^+}^{\text{FEP}} = -21.3 \pm 0.4$  kcal/mol from the hydration free energies computed directly using FEP on the individual ions (Table 1, step 3-3), confirming that the mutations are independent.

The hydration free energy of the neutral salt, computed from concurrent dual mutations,  $\Delta G_1^{\text{FEP}} = \Delta G_{N^0 \rightarrow Cl^-} + \Delta G_{N^0 \rightarrow K^+} = -162.8 \pm 0.7$  kcal/mol is in excellent agreement with that computed using separate FEP computations,  $\Delta G_1^{\text{FEP}} = \Delta G_{N^0 \rightarrow Cl^-}^{\text{FEP}} + \Delta G_{N^0 \rightarrow K^+}^{\text{FEP}} = -162.5 \pm 0.4$  kcal/mol. These hydration free energies agree with the results for  $\Delta G_{N^0 \rightarrow K^+}$  and  $\Delta G_{N^0 \rightarrow Cl^-}$  from previous work using the CHARMM 27 force field and the TIP3P water model.<sup>42,57,61</sup>

The good agreement between the computed and experimental values for the hydration free energy of the neutral salt (potassium chloride) of  $-160.4$  kcal/mol by Tissandier et al.<sup>53</sup> (corrected for the typographical errors detailed by Kelly et al.<sup>65</sup>) is a measure of the quality of, and compromises inherent in, the ion parametrization with the TIP3P model; any discrepancy can be attributed to the limitations of the model and/or parametrization. Although the concurrent dual mutation method is intended for applications where theoretical treatments of the self-energy correction are not as easily applied, such as mutations in biomolecules, this exercise serves as proof-of-concept.

2. *Effect of Radius Dependence on Free Energy Change.* Analytical self-energy corrections for infinitely periodic systems also include a dependence on solvent permittivity and solute radius.<sup>21,26,58</sup> The radius-dependent correction is intended to account for finite solute size and finite solvent permittivity in a periodic system.<sup>21,58</sup> It is defined as

$$\Delta \Delta G_{\text{solv}} = -\frac{1}{24\pi\epsilon_0 L} (e_i^{-1} - \epsilon_s^{-1}) \left\{ \xi'_{\text{EW}} + \frac{4\pi}{3} \left( \frac{R}{L} \right)^2 - \frac{16\pi^2}{45} \left( \frac{R}{L} \right)^3 \right\} \text{ for } R \leq \frac{1}{2}L \quad (21)$$

Given  $\xi'_{\text{EW}} = -2.837$  and  $\epsilon_i = 1$  (internal permittivity), in the limit of  $\epsilon_s \rightarrow \infty$  (solvent permittivity) and  $R \rightarrow 0$  (hard-sphere radius),  $\Delta \Delta G_{\text{solv}} \rightarrow C_{\text{EW}}$ . The radius for atomic ions is defined by the excluded volume, which can be approximated as the Lennard-Jones radius or calculated empirically

cally from radial distribution functions.<sup>57</sup> Solvent permittivity is a constant for a given water model.

The radius dependence in eq 21 could pose an issue for the cancellation of  $E_{\text{self}}$  in cases where the sizes of the two mutations are dissimilar, such as an atomic ion and a residue within a protein. As a test, step 1 was applied to the mutation  $\text{D}^- + \text{N}^0 \rightarrow \text{S}^0 + \text{K}^+$ , where the mutation of aspartic acid (D,  $q = -1$ ) to serine (S,  $q = 0$ ) is denoted as  $\text{D}^- \rightarrow \text{S}^0$  with free energy change  $\Delta G_{\text{D}^- \rightarrow \text{S}^0}$ . Note that the direction of the change in charge is reversed for the negative species from the example ( $\text{N}^0 \rightarrow \text{Cl}^-$ ), which changes the procedure slightly. The mutation  $\text{N}^0 \rightarrow \text{K}^+$  has the free energy change  $\Delta G_{\text{N}^0 \rightarrow \text{K}^+}$ , as defined previously. The zwitterionic forms of  $\text{D}^-$  and  $\text{S}^0$  were used, and only the position of  $\text{K}^+$  was fixed in the simulation box ( $L = 40 \text{ \AA}$ ). The results of each step are listed in Table 2.

We find  $\Delta G_{\text{D}^- \rightarrow \text{S}^0} = 153.4 \pm 1.5 \text{ kcal/mol}$  and  $\Delta G_{\text{N}^0 \rightarrow \text{K}^+} = -70.9 \pm 1.5 \text{ kcal/mol}$  (Table 2, step 1–3).  $\Delta G_1^{\text{FEP}} = 82.5 \pm 1.3 \text{ kcal/mol}$  of the concurrent dual mutation (Table 2, step 1–1) is in good agreement with  $\Delta G_{\text{D}^- \rightarrow \text{S}^0}^{\text{FEP}} + \Delta G_{\text{N}^0 \rightarrow \text{K}^+}^{\text{FEP}} = 81.8 \pm 0.7 \text{ kcal/mol}$  from the combined single mutations (Table 2, step 3–3); therefore the mutations are independent. To calculate  $\Delta G_{\text{D}^- \rightarrow \text{S}^0}$  using step 2, subtract  $\Delta G_{\text{N}^0 \rightarrow \text{K}^+}$  (Table 1, step 1–3) from  $\Delta G_1^{\text{FEP}}$  (Table 2, step 1–1) to obtain  $\Delta G_{\text{D}^- \rightarrow \text{S}^0} = 153.2 \pm 2.6 \text{ kcal/mol}$ . This result is in agreement with step 1–3, where  $\Delta G_{\text{D}^- \rightarrow \text{S}^0} = 153.4 \pm 1.5 \text{ kcal/mol}$  (Table 2, step 1–3).

$\Delta G_{\text{N}^0 \rightarrow \text{K}^+}$  obtained from simulations where the second mutation is either an atomic ion of similar radius ( $\text{Cl}^-$ ) or a larger amino acid are nearly identical ( $-70.7 \pm 1.3$  versus  $-70.9 \pm 1.5 \text{ kcal/mol}$ , respectively). The self-energy contributions to  $\Delta G_{\text{N}^0 \rightarrow \text{Cl}^-}$  and  $\Delta G_{\text{D}^- \rightarrow \text{S}^0}$  must be the same to produce such similar results; therefore, the self-energy depends on neither the radius nor the identity of the ionic species, to within the error. This result highlights a strength of the method presented in this work.  $\Delta \Delta G_{\text{solv}}$  depends on  $R/L$ ; hence, the effect of the difference in the radii should be minimal for sufficiently large  $L$ . However, the radius of a complex molecule is not usually well-defined. In addition, eq 21 was derived for a hard sphere, as is appropriate for atomic ions, but the accuracy of the correction for other solutes is unknown. Fortunately, the self-energy cancels implicitly in this method, and the radius is not needed to calculate the free energy change.

**3. Comparison of  $E_{\text{self}}$  to Correction Terms.** The FEP of the single mutations calculated for step 3 yields  $\Delta G + E_{\text{self}}$ . Having previously obtained the free energy change, we can calculate the self-energy and compare it to theoretical corrections such as  $C_{\text{EW}}$  and  $\Delta \Delta G_{\text{solv}}$  (Table 3).

The self-energy was calculated for  $\text{N}^0 \rightarrow \text{K}^+$  and  $\text{N}^0 \rightarrow \text{Cl}^-$  in system sizes of  $L = 30, 20,$  and  $11 \text{ \AA}$  (exact values are given in Table 3).  $E_{\text{self}}$  was calculated by subtracting the free energy changes (Table 1) from  $\Delta G^{\text{FEP}}$  of the single ion mutations.  $C_{\text{EW}}$  is defined by eq 10, and  $\Delta \Delta G_{\text{solv}}$  was calculated from eq 21 using  $\epsilon_s = 78$  and  $R = 3.53$  and  $4.54 \text{ \AA}$  for  $\text{K}^+$  and  $\text{Cl}^-$ , respectively, as defined by the CHARMM 27 force field.<sup>64</sup>

The system sizes are not identical as the FEP calculations are performed in the isobaric–isothermal ensemble; however,

**Table 4.** Evaluation of System Size Dependence via Free Energy Changes (kcal/mol)<sup>a</sup>

$L$ (Å)	$\Delta G_{\text{N}^0 \rightarrow \text{Cl}^-}^{\text{FEP}} + C_{\text{EW}}^b$	$\Delta G_{\text{N}^0 \rightarrow \text{Cl}^-} + \Delta \Delta G_{\text{solv}}^c$	$\Delta G_{\text{N}^0 \rightarrow \text{Cl}^-}^{\text{FEP}} - \Delta G_{\text{N}^0 \rightarrow \text{K}^+}^{\text{FEP}}^d$
29.11	$-108.1 \pm 0.3$	$-92.7 \pm 0.3$	$-21.3 \pm 0.4$
18.97	$-115.6 \pm 0.1$	$-93.2 \pm 0.1$	$-20.9 \pm 0.3$
10.78	$-131.7 \pm 0.2$	$-99.2 \pm 0.2$	$-22.6 \pm 0.4$

<sup>a</sup> All values are derived from Table 3. The errors are from the free energy calculated via FEP ( $\Delta G^{\text{FEP}}$ ) for a neutral dummy atom ( $\text{N}^0$ ) into a potassium ion ( $\text{K}^+$ ) or chloride ion ( $\text{Cl}^-$ ) for three system sizes ( $L$ ). <sup>b</sup>  $\Delta G_{\text{N}^0 \rightarrow \text{Cl}^-}^{\text{FEP}} + C_{\text{EW}}$  is the final free energy change as estimated by the theoretical correction terms from eq 10.<sup>41</sup> No contribution of  $C_{\text{EW}}$  and  $\Delta \Delta G_{\text{solv}}$  to the error is included. <sup>c</sup>  $\Delta G_{\text{N}^0 \rightarrow \text{Cl}^-} + \Delta \Delta G_{\text{solv}}$  is the final free energy change as estimated by the theoretical correction terms from and eq 21.<sup>21</sup> No contribution of  $C_{\text{EW}}$  and  $\Delta \Delta G_{\text{solv}}$  to the error is included. <sup>d</sup>  $\Delta G_{\text{N}^0 \rightarrow \text{Cl}^-}^{\text{FEP}} - \Delta G_{\text{N}^0 \rightarrow \text{K}^+}^{\text{FEP}}$  is step 3–3 from Table 1 recalculated using  $\Delta G^{\text{FEP}}$  from each system size.

$E_{\text{self}}$  and the theoretical correction terms do not appear to be overly sensitive to minor differences in the system size. The self-energy terms of the two ions are comparable for each system size, confirming that  $E_{\text{self}}$  does cancel.

$C_{\text{EW}}$  greatly overestimates the value of the self-energy term. For  $L = 30 \text{ \AA}$ ,  $E_{\text{self}} = 0.1$  and  $0.3 \text{ kcal/mol}$  for  $\text{N}^0 \rightarrow \text{K}^+$  and  $\text{N}^0 \rightarrow \text{Cl}^-$ , respectively, while  $C_{\text{EW}} = -16.2 \text{ kcal/mol}$  for both ions (Table 3). This discrepancy worsens as the system size decreases.  $\Delta \Delta G_{\text{solv}}$  is slightly higher than  $E_{\text{self}}$ , but within the error, except for  $\text{N}^0 \rightarrow \text{Cl}^-$  when  $L = 11 \text{ \AA}$  (Table 3). The similarity of  $E_{\text{self}}$  and  $\Delta \Delta G_{\text{solv}}$  indicates that additional dependencies are likely small contributions to the self-energy. At  $L = 30 \text{ \AA}$ , we consider  $E_{\text{self}}$  to be a negligible contribution to the free energy change for these mutations.

The relative effect of these corrections on the final free energy change can be seen by adding  $C_{\text{EW}}$  or  $\Delta \Delta G_{\text{solv}}$  to  $\Delta G^{\text{FEP}}$  (Table 4). For  $\Delta G_{\text{N}^0 \rightarrow \text{Cl}^-}^{\text{FEP}}$  with  $L = 30 \text{ \AA}$ , the final free energy changes are  $\Delta G_{\text{N}^0 \rightarrow \text{Cl}^-} = -108.1$  and  $-92.7 \text{ kcal/mol}$  for  $\Delta G_{\text{N}^0 \rightarrow \text{Cl}^-}^{\text{FEP}} + C_{\text{EW}}$  and  $\Delta G_{\text{N}^0 \rightarrow \text{Cl}^-}^{\text{FEP}} + \Delta \Delta G_{\text{solv}}$ , respectively. Table 4 shows a system size dependence in  $\Delta G_{\text{N}^0 \rightarrow \text{Cl}^-}$  calculated using the theoretical correction terms  $C_{\text{EW}}$  and  $\Delta \Delta G_{\text{solv}}$ . For  $L = 11 \text{ \AA}$ ,  $\Delta G_{\text{N}^0 \rightarrow \text{Cl}^-} = -131.7$  and  $-99.2 \text{ kcal/mol}$  for  $\Delta G_{\text{N}^0 \rightarrow \text{Cl}^-}^{\text{FEP}} + C_{\text{EW}}$  and  $\Delta G_{\text{N}^0 \rightarrow \text{Cl}^-}^{\text{FEP}} + \Delta \Delta G_{\text{solv}}$ , respectively. Residual dependence on system size is an indication of inaccuracy in these correction terms.

The system size dependence of the method presented here can be evaluated by recalculating step 3–3 in Table 1 using different system sizes. In this step, the independence of the mutations is evaluated by comparing the difference in the free energy changes calculated either concurrently or individually. Above, we found  $\Delta G_2^{\text{FEP}} = -21.5 \pm 0.6 \text{ kcal/mol}$  for the concurrent dual mutation and  $\Delta G_{\text{N}^0 \rightarrow \text{Cl}^-}^{\text{FEP}} - \Delta G_{\text{N}^0 \rightarrow \text{K}^+}^{\text{FEP}} = -21.3 \pm 0.4 \text{ kcal/mol}$  for the combined single mutations ( $L = 30 \text{ \AA}$ ). If instead we calculate  $\Delta G_{\text{N}^0 \rightarrow \text{Cl}^-}^{\text{FEP}} - \Delta G_{\text{N}^0 \rightarrow \text{K}^+}^{\text{FEP}}$  using  $\Delta G^{\text{FEP}}$  of  $L = 20 \text{ \AA}$  and  $11 \text{ \AA}$ , we find  $\Delta G_{\text{N}^0 \rightarrow \text{Cl}^-}^{\text{FEP}} - \Delta G_{\text{N}^0 \rightarrow \text{K}^+}^{\text{FEP}} = -20.9 \pm 0.3$  and  $-22.6 \pm 0.4 \text{ kcal/mol}$ , respectively (Table 4). The cancellation of the self-energy in  $\Delta G_{\text{N}^0 \rightarrow \text{Cl}^-}^{\text{FEP}} - \Delta G_{\text{N}^0 \rightarrow \text{K}^+}^{\text{FEP}}$  requires that the two system sizes be equal. From Table 3, it is evident that this is not always the case in the isothermal–isobaric ensemble, which is an additional source of error not included in Table



4. For large system sizes, this error is minimal, but for  $L = 11 \text{ \AA}$ , the difference is likely not negligible. Nevertheless, there is very little dependence on system size in the results from the method presented in this paper.

4. *Efficiency.* The efficiency of the method presented in this work is similar to current calculations for neutral mutations, as FEP of atomic ions (step 1) is computationally inexpensive compared to FEP in a biomolecule required for any method (step 2). If parallel computing resources are available, this method adds little additional wall-clock time. All of the simulations, including the optional calculations for verifying independence (step 3), can be run in parallel.

In FEP calculations performed with biomolecules, the system sizes are often much larger than those used here. The two single mutation FEP simulations must be performed using the same system size, which should be chosen to minimize nonbonded interactions between the larger solute and its periodic images. However, since the verification step (step 3) is independent of system size, the single mutation FEP simulations are not required to use the same system size as the dual mutation FEP simulation. This could be of particular use if interactions between the two mutations are an issue; if the dual mutation FEP needs to be rerun using a larger system size, the smaller single mutation FEP simulations can still be used to check for independence. In general, smaller system sizes can be used for the single mutation FEP simulations to make the verification of independence less computationally costly, provided that the system size is reasonable.

The ability to divide large charge mutations into smaller steps allows the staging methods often implemented for  $\lambda$  to improve accuracy<sup>62</sup> to be applied to  $q$  as well. While this does not increase the computational efficiency of the final calculation, if  $q$  is not linear in  $\lambda$ , it may be easier to find an optimal staging scheme for  $q$  than for  $\lambda$ .

We have shown that the method presented in this work allows the free energy change for charge mutations to be determined without the limitations of current approaches. The mutations were confirmed as independent in the cases demonstrated here, validating the use of concurrent dual mutations. We find that the self-energy is primarily dependent on charge; other dependences in the self-energy, such as radius, are seen to have little affect. We also find that, at  $L = 30 \text{ \AA}$ , the system is sufficiently large to minimize the self-interactions for these mutations.

## VI. Summary and Conclusions

The dependence of the self-energy on both charge and system size has been a barrier to accurately calculating the free energy change for charge mutations by FEP or TI using the Ewald summation for electrostatics. The method presented here removes the need for precise knowledge of the self-energy contribution to the free energy introduced by the periodicity in the Ewald implementation.

The method presented in this work translates the results from studies of charge mutations in atomic ions to more complex systems. The solvation free energy of atomic ions can be accurately obtained, but it requires analytical corrections which are not well-defined, or well studied, for

multiatom mutations. The method presented in this work is general in its implementation but will be most useful for calculations of binding free energy differences,  $\Delta\Delta G$ , of biomolecules, as illustrated in Figure 1. Unlike current methods used for charge mutations in biomolecules,<sup>10–12,49</sup> it does not require experimental data to confirm the accuracy of the self-energy correction. It is self-contained and requires only self-consistency in the simulation parameters. Moreover, the assumptions made here can be tested for any system.

The uncertainty of the self-energy correction is removed because the self-energy terms implicitly cancel with the method presented here. Furthermore, the self-energy can be determined using the free energy change, and self-energy dependencies can be tested for any complex system. The self-energy can provide a quantitative metric for the extent of self-interactions in MD simulations that use the Ewald implementation.

Electrostatic interactions are the source of all nonbonded interactions in the physical world. The ability to calculate the free energy of perturbation associated with a charge differential without the uncertainty associated with large corrections has many important applications. Mutation studies can be done computationally for any residue, regardless of charge. The effect of ions on DNA, RNA, and ion channels can be directly studied, and the effect of charge on ligand binding can be isolated. The ability to decompose a thermodynamic reaction into intermediate steps allows small increments of charge to be used, which enables more general studies to explore the effect of electrostatic interactions on protein dynamics and binding as a function of charge.

## References

- (1) Beveridge, D. L.; Dicapua, F. M. *Annu. Rev. Biophys. Chem.* **1989**, *18*, 431–492.
- (2) Straatsma, T. P.; McCammon, J. A. *Annu. Rev. Phys. Chem.* **1992**, *43*, 407–435.
- (3) Chipot, C.; Pearlman, D. A. *Mol. Simul.* **2002**, *28*, 1–12.
- (4) Bash, P. A.; Singh, U. C.; Langridge, R.; Kollman, P. A. *Science* **1987**, *236*, 564–568.
- (5) Dixit, S. B.; Bhasin, R.; Rajasekaran, E.; Jayaram, B. *J. Chem. Soc., Faraday Trans.* **1997**, *93*, 1105–1113.
- (6) Shirts, M. R.; Pande, V. S. *J. Chem. Phys.* **2005**, *122*, 134508–134508–13.
- (7) Bren, U.; Martinek, V.; Florian, J. *J. Phys. Chem. B* **2006**, *110*, 12782–12788.
- (8) Deng, Y. Q.; Roux, B. *J. Phys. Chem. B* **2004**, *108*, 16567–16576.
- (9) Shivakumar, D.; Deng, Y. Q.; Roux, B. *J. Chem. Theory Comput.* **2009**, *5*, 919–930.
- (10) Zhou, R. H.; Das, P.; Royyuru, A. K. *J. Phys. Chem. B* **2008**, *112*, 15813–15820.
- (11) Donnini, S.; Mark, A. E.; Juffer, A. H.; Villa, A. *J. Comput. Chem.* **2005**, *26*, 115–122.
- (12) Das, P.; Li, J. Y.; Royyuru, A. K.; Zhou, R. H. *J. Comput. Chem.* **2009**, *30*, 1654–1663.
- (13) Michielin, O.; Karplus, M. *J. Mol. Biol.* **2002**, *324*, 547–569.



- (14) Dixit, S. B.; Chipot, C. *J. Phys. Chem. A* **2001**, *105*, 9795–9799.
- (15) Pan, Y. M.; Gao, D. Q.; Yang, W. C.; Cho, H.; Zhan, C. G. *J. Am. Chem. Soc.* **2007**, *129*, 13537–13543.
- (16) Deng, Y. Q.; Roux, B. *J. Phys. Chem. B* **2009**, *113*, 2234–2246.
- (17) Frenkel, D.; Smit, B. Free energy calculations. In *Understanding molecular simulation: from algorithms to applications*, 2nd ed.; Elsevier: San Diego, CA, 1996; pp 167–171.
- (18) Allen, M. P.; Tildesley, D. J. Free energy estimation. In *Computer simulation of liquids*; Oxford University Press: New York, 1987; pp 213–219.
- (19) Gao, J.; Kuczera, K.; Tidor, B.; Karplus, M. *Science* **1989**, *244*, 1069–1072.
- (20) Kollman, P. *Chem. Rev.* **1993**, *93*, 2395–2417.
- (21) Hunenberger, P. H.; McCammon, J. A. *J. Chem. Phys.* **1999**, *110*, 1856–1872.
- (22) Figueirido, F.; Delbuono, G. S.; Levy, R. M. *J. Chem. Phys.* **1995**, *103*, 6133–6142.
- (23) Sakane, S.; Ashbaugh, H. S.; Wood, R. H. *J. Phys. Chem. B* **1998**, *102*, 5673–5682.
- (24) Bogusz, S.; Cheatham, T. E.; Brooks, B. R. *J. Chem. Phys.* **1998**, *108*, 7070–7084.
- (25) Kastenzholz, M. A.; Hunenberger, P. H. *J. Chem. Phys.* **2006**, *124*, 124106–124106–27.
- (26) Kastenzholz, M. A.; Hunenberger, P. H. *J. Chem. Phys.* **2006**, *124*, 224501–224501–20.
- (27) Aqvist, J.; Hansson, T. *J. Phys. Chem. B* **1998**, *102*, 3837–3840.
- (28) Aqvist, J. *J. Comput. Chem.* **1996**, *17*, 1587–1597.
- (29) Carlsson, J.; Aqvist, J. *J. Phys. Chem. B* **2009**, *113*, 10255–10260.
- (30) Carlsson, J.; Aqvist, J. *Phys. Chem. Chem. Phys.* **2006**, *8*, 5385–5395.
- (31) Straatsma, T. P.; Berendsen, H. J. C. *J. Chem. Phys.* **1988**, *89*, 5876–5886.
- (32) Aqvist, J. *J. Phys. Chem.* **1990**, *94*, 8021–8024.
- (33) Haranczyk, M.; Gutowski, M.; Warshel, A. *Phys. Chem. Chem. Phys.* **2008**, *10*, 4442–4448.
- (34) Jorgensen, W.; Ulmschneider, J.; Tirado-Rives, J. *J. Phys. Chem. B* **2004**, *108*, 16264–17270.
- (35) Almlof, M.; Carlsson, J.; Aqvist, J. *J. Chem. Theory Comput.* **2007**, *3*, 2162–2175.
- (36) Harder, E.; Roux, B. *J. Chem. Phys.* **2008**, *129*, 234706–234706–9.
- (37) Lamoureux, G.; Roux, B. *J. Phys. Chem. B* **2006**, *110*, 3308–3322.
- (38) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.
- (39) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.
- (40) Petersen, H. G. *J. Chem. Phys.* **1995**, *103*, 3668–3679.
- (41) Hummer, G.; Pratt, L. R.; Garcia, A. E. *J. Phys. Chem.* **1996**, *100*, 1206–1215.
- (42) Grossfield, A.; Ren, P. Y.; Ponder, J. W. *J. Am. Chem. Soc.* **2003**, *125*, 15671–15682.
- (43) Figueirido, F.; DelBuono, G. S.; Levy, R. M. *J. Phys. Chem. B* **1997**, *101*, 5622–5623.
- (44) Rozanska, X.; Chipot, C. *J. Chem. Phys.* **2000**, *112*, 9691–9694.
- (45) Marrone, T. J.; Merz, K. M. *J. Phys. Chem.* **1993**, *97*, 6524–6529.
- (46) Deleueuw, S. W.; Perram, J. W.; Smith, E. R. *Annu. Rev. Phys. Chem.* **1986**, *37*, 245–270.
- (47) Cichocki, B.; Felderhof, B. U.; Hinsen, K. *Phys. Rev. A* **1989**, *39*, 5350–5358.
- (48) Nijboer, B. R. A.; Ruijgrok, T. W. *J. Stat. Phys.* **1988**, *53*, 361–382.
- (49) Nina, M.; Beglov, D.; Roux, B. *J. Phys. Chem. B* **1997**, *101*, 5239–5248.
- (50) Gomer, R.; Tryson, G. *J. Chem. Phys.* **1977**, *66*, 4413–4424.
- (51) Klots, C. E. *J. Phys. Chem.* **1981**, *85*, 3585–3588.
- (52) Schmid, R.; Miah, A. M.; Sapunov, V. N. *Phys. Chem. Chem. Phys.* **2000**, *2*, 97–102.
- (53) 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.
- (54) Noyes, R. M. *J. Am. Chem. Soc.* **1962**, *84*, 513–522.
- (55) Marcus, Y. *J. Chem. Soc., Faraday Trans.* **1991**, *87*, 2995–2999.
- (56) Randles, J. E. B. *J. Trans. Faraday Soc.* **1956**, *52*, 1573–1581.
- (57) Warren, G. L.; Patel, S. *J. Chem. Phys.* **2007**, *127*, 064509–064509–19.
- (58) Hummer, G.; Pratt, L. R.; Garcia, A. E. *J. Chem. Phys.* **1997**, *107*, 9275–9277.
- (59) *Free energy calculations*; Chipot, C., Pohorille, A., Eds.; Springer: New York, 2007.
- (60) Zwanikken, J.; van Roij, R. *J. Phys.: Condens. Matter* **2009**, *21*, 424102–424102–6.
- (61) Rajamani, S.; Ghosh, T.; Garde, S. *J. Chem. Phys.* **2004**, *120*, 4457–4466.
- (62) Lu, N.; Kofke, D. A.; Woolf, T. B. *J. Phys. Chem. B* **2003**, *107*, 5598–5611.
- (63) 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.
- (64) 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.
- (65) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2006**, *110*, 16066–16081.

## Magnetic Exchange Couplings with Range-Separated Hybrid Density Functionals

 Juan E. Peralta\*<sup>†</sup> and Juan I. Melo<sup>‡</sup>

*Department of Physics, Central Michigan University, Mt. Pleasant, Michigan 48859, and Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Cdad. Universitaria, Pab. I, 1428 Buenos Aires, Argentina and CONICET*

Received February 22, 2010

**Abstract:** We investigate the effect of Hartree–Fock range-separation on the calculation of magnetic exchange couplings in a set of nine bimetallic transition-metal complexes containing 3d elements (V, Cr, Mn, and Cu). To this end, we have compared magnetic exchange couplings calculated as self-consistent energy differences using two global hybrid functionals, B3LYP (Becke 3-parameter exchange and Lee–Yang–Parr correlation) and PBEh (hybrid Perdew–Burke–Ernzerhof) with the short-range separated HSE (Heyd–Scuseria–Ernzerhof) and the long-range corrected LC- $\omega$ PBE. Our results show that, although there is no clear superiority of any of these functionals when compared with experimental data, the LC- $\omega$ PBE provides a better description of the magnetization on the metallic centers, yielding self-consistent solutions that mimic more closely a Heisenberg-like behavior.

### Introduction

One of the most interesting properties of molecular complexes containing transition-metal atoms is their ability to behave as single-molecule magnets. Many applications have been proposed exploiting these molecular-size magnets, such as quantum computation units and high-density data storage.<sup>1</sup> The magnetic behavior of these systems has been probed using a variety of experimental techniques. In all cases, an empirical model based on the Heisenberg spin Hamiltonian was found to fit the experimental data very well, provided that the parameters in the model Hamiltonian are chosen properly. Thus, modeling the magnetism of single-molecule magnets can be reduced to analyzing simple statistical models based on a Heisenberg spin Hamiltonian that includes both external parameters (temperature, applied magnetic field, etc.) and internal parameters (magnetic exchange couplings, magnetic anisotropy, etc.).

Internal parameters for a particular molecular magnet can be obtained from first-principles electronic structure calcula-

tions by mapping the molecular energies to the energies of the Heisenberg spin Hamiltonian.<sup>2–4</sup> In particular, magnetic exchange couplings,  $J$ , can be obtained considering the isotropic Heisenberg Hamiltonian

$$H = -2 \sum_{\langle i,j \rangle} J_{ij} \mathbf{S}_i \cdot \mathbf{S}_j \quad (1)$$

where  $\mathbf{S}_i$  and  $\mathbf{S}_j$  are the (localized) spin operators associated to each magnetic center. The theoretical prediction of magnetic exchange couplings depends mainly on two factors: the approach employed in this mapping and the choice of the electronic structure method. Because of the size of most systems of interest, density functional theory (DFT) represents the most viable electronic structure method to this end.

Several approaches have been proposed for extracting  $J$  couplings from DFT energies. According to the spin-projected (SP) approach,<sup>2</sup> the energies of a two-center complex  $A$  and  $B$  can be related to the  $J$  coupling as

$$E_{LS} - E_{HS} = 4S_A S_B J_{AB} \quad (2)$$

while in the non(spin)projected (NP) approach,<sup>5</sup> the energies of a two-center complex  $S_A$  and  $S_B$  can be related to the  $J$  coupling as

\* To whom correspondence should be addressed. E-mail: juan.peralta@cmich.edu.

<sup>†</sup> Department of Physics, Central Michigan University.

<sup>‡</sup> Departamento de Física, Universidad de Buenos Aires.

$$E_{\text{LS}} - E_{\text{HS}} = (4S_A S_B + 2S_B) J_{AB} \quad (3)$$

where  $S_B \leq S_A$ . In eqs 2 and 3,  $E_{\text{HS}}$  is the energy of the high-spin state and  $E_{\text{LS}}$  is the energy of the low-spin (broken-symmetry) state. Equations 2 and 3 can be straightforwardly generalized to a set of equations for complexes with multiple magnetic centers.<sup>6,7</sup> While the SP and NP methods are fairly popular, other methods have been proposed in the literature such as Nishino's approach<sup>8</sup> and the constrained-DFT approach of Rudra et al.<sup>9,10</sup>

It has been shown that for the calculation of magnetic exchange couplings, hybrid functionals perform the best among several realizations of density functionals available in the literature.<sup>3</sup> In particular, Ruiz et al. have shown that the broken-symmetry approach in combination with the hybrid B3LYP functional<sup>11,12</sup> yields the best exchange couplings between several popular density functionals when compared to experimental values.<sup>5</sup> Valero et al. have recently shown that the M06 realization of the generalized-gradient approximation (GGA) functional yields exchange couplings as accurate as B3LYP.<sup>13</sup> It has been suggested that the presence of self-interaction error (SIE) in approximate density functionals mimics in some way nondynamical electron correlation contributions to the calculated energies.<sup>14,15</sup> Since the use of the NP approximation also accounts for electron correlation that is not included in the spin-projected approximation, it was argued by Ruiz et al. that using the broken-symmetry method in combination with a self-interaction free functional should give accurate exchange couplings,<sup>16</sup> although this led to some controversy.<sup>17–19</sup> Several authors have argued that eq 2 represents a more physically meaningful mapping between the Heisenberg and DFT models<sup>20–24</sup> than eq 3 and that the accuracy of B3LYP combined with eq 3 is fortuitous. Therefore, a density functional that is able to reproduce magnetic exchange couplings in combination with eq 2 would be desirable.

A new generation of density functionals that incorporate screened Hartree–Fock (HF) exchange became recently available. Such is the case of the Heyd–Scuseria–Ernzerhof (HSE) functional,<sup>25–27</sup> which includes a portion of short-range HF exchange in its definition that makes it suitable to treat electronic localization effects and, at the same time, computationally more efficient than traditional (global) hybrids. The LC- $\omega$ PBE,<sup>28</sup> which incorporates long-range HF exchange to partly remove SIE, provides a not exactly one-electron, but most often “many-electron self interaction-free” functional.<sup>29</sup> Rivero et al. have analyzed the reliability of these range-separated hybrid functionals for describing magnetic exchange interactions using a reference database proposed by Valero et al.<sup>13</sup> In view of these developments, it is important to investigate the performance of these new models for the prediction of magnetic parameters. It is the purpose of this work to compare magnetic exchange couplings calculated with the range-separated HSE and LC- $\omega$ PBE with those calculated with global hybrid functionals.

## Methodology

All magnetic exchange couplings were calculated from self-consistent field (SCF) energy differences for the HS and LS

states, as given by eqs 2 and 3 for the SP and NP approaches, respectively. The Gaussian Development Version was used through this work.<sup>30</sup> The low-spin solution was obtained from an initial SCF guess generated by flipping the local spin-density in one of the metal centers of the high-spin solution. We have verified that the SCF solutions approximately represent the target Heisenberg solutions by comparing Mulliken atomic spin densities for each particular case. All calculations converged the SCF procedure to an accuracy of  $10^{-8}$  hartree = 0.27  $\mu\text{eV}$  in the total energy. An atom-centered numerical integration grid of 99 radial and 590 angular points (grid = ultrafine keyword in Gaussian) was employed in all cases. Geometrical structures were taken from experimental crystallographic data. All calculations were carried out using the Ahlrich's double- $\zeta$  valence plus polarization Gaussian basis for atoms other than transition metals<sup>31</sup> and all-electron Ahlrich's triple- $\zeta$  valence plus polarization for the metal centers.<sup>32</sup> Molecular data (spin configurations, total charge, and spin multiplicities), geometrical structures, and complete basis sets are available as Supporting Information.

To assess the effect of range separation in density functionals we have chosen a set of nine bimetallic transition-metal complexes containing 3d elements (V, Cr, Mn, and Cu). Five of them (compounds 1–5) present antiferromagnetically coupled magnetic centers ( $J_{AB} < 0$ ), while the remaining four (compounds 6–9) are ferromagnetically coupled ( $J_{AB} > 0$ ). These systems have been employed by Rudra et al. to assess the performance of a proposed methodology to evaluate magnetic exchange couplings based on constraint DFT.<sup>9</sup> Here we have evaluated magnetic exchange couplings for these 9 complexes using the global hybrid functionals B3LYP and PBEh,<sup>33–35</sup> and the range-separated hybrid functionals HSE and LC- $\omega$ PBE.

## Results and Discussion

In Table 1, we show our results for the magnetic exchange couplings. Experimental values and their corresponding reference are also shown. For all four hybrid functionals, exchange couplings evaluated using the NP approach are in slightly closer agreement with the available experimental data. Our NP and SP results for the B3LYP functional are in line with those of Rudra et al.<sup>9</sup> (not shown in Table 1), although the calculated couplings are somewhat different. We attribute this discrepancy to the different basis set employed by Rudra et al. and this work.

For all antiferromagnetically coupled complexes (compounds 1–5), the PBEh functional yields a weaker coupling between magnetic centers as compared to B3LYP, while the trend for ferromagnetically coupled complexes is not uniform. On the other hand, the effect of truncating the long-range HF exchange in the HSE functional effectively reduces ferromagnetic exchange couplings (compounds 6–9) and increases antiferromagnetic couplings (compounds 1–5), as evidenced by comparing HSE and PBEh results. Contrarily, the LC- $\omega$ PBE functional produces the opposite effect in most cases, with the exception of the  $\text{Mn}^{\text{III}}\text{Mn}^{\text{IV}}$  (compound 5) and  $\text{Cu}^{\text{II}}\text{Cr}^{\text{III}}$  (compound 8) complexes. In particular, PBEh and HSE show a very large deviation for the latter complex

**Table 1.** Magnetic Exchange Couplings (in  $\text{cm}^{-1}$ ) Calculated with Different Hybrid Density Functionals<sup>a</sup>

complex	B3LYP		PBEh		HSE		LC- $\omega$ PBE		ref 9	
	SP	NP	SP	NP	SP	NP	SP	NP	C-DFT	exptl
(1) Cu <sup>II</sup> -Cu <sup>II</sup>	-84.2	-42.1	-59.5	-29.8	-63.5	-31.8	-40.3	-20.1	-16	-30.9 <sup>b</sup>
(2) Cu <sup>II</sup> -Cu <sup>II</sup>	-101.8	-50.9	-79.0	-39.5	-83.4	-41.7	-58.2	-29.1	-44	-37.4 <sup>c</sup>
(3) Mn <sup>II</sup> -Cu <sup>II</sup>	-36.4	-30.4	-27.8	-23.2	-29.7	-24.8	-18.5	-15.4	-128	-15.7 <sup>d</sup>
(4) V <sup>IV</sup> -V <sup>IV</sup>	-100.2	-50.1	-82.3	-41.1	-87.4	-43.7	-62.4	-31.2	-83	-107 <sup>e</sup>
(5) Mn <sup>III</sup> Mn <sup>IV</sup>	-171.0	-136.8	-138.0	-110.4	-142.5	-114.0	-152.7	-122.1	-128	-110 <sup>f</sup>
(6) Cu <sup>II</sup> -Cu <sup>II</sup>	103.5	51.7	140.6	70.3	133.8	66.9	247.7	123.8	112	84 <sup>g</sup>
(7) Cu <sup>II</sup> -Cu <sup>II</sup>	131.9	66.0	119.9	60.0	119.2	59.6	120.2	60.1	57	57 <sup>h</sup>
(8) Cu <sup>II</sup> Cr <sup>III</sup>	14.6	11.0	170.9	128.2	169.1	126.8	8.1	6.1	23	18.5 <sup>i</sup>
(9) Cu <sup>II</sup> Mn <sup>III</sup>	75.6	60.5	28.8	23.1	10.8	8.7	46.9	37.5	75	54.4 <sup>j</sup>
MAE	36.2	19.8	48.1	26.1	50.1	28.4	40.6	19.9	25.4	
MAE excluding (8)	40.2	21.3	35.0	15.6	37.5	18.4	44.3	20.9		

<sup>a</sup> MAE indicates the mean absolute error compared with experimental data. Magnetic exchange couplings taken from ref 9 are based on constraint-DFT (C-DFT) calculations. <sup>b</sup> Taken from ref 43. <sup>c</sup> Taken from ref 44. <sup>d</sup> Taken from ref 45. <sup>e</sup> Taken from ref 46. <sup>f</sup> Taken from ref 47. <sup>g</sup> Taken from ref 48. <sup>h</sup> Taken from ref 49. <sup>i</sup> Taken from ref 50.

compared with the B3LYP and LC- $\omega$ PBE functionals and experimental results. In most cases, HSE and PBEh exchange couplings differ a few  $\text{cm}^{-1}$ , with the exception of compound 9 where the difference is about  $14 \text{ cm}^{-1}$ . This implies that the truncation of the long-range HF exchange in the HSE functional has little impact on the calculated magnetic exchange couplings.

Overall, B3LYP and LC- $\omega$ PBE provide a very good agreement with experimental values when the NP approximation is employed to map the DFT energies to the model Hamiltonian energies. This has been noted by different authors,<sup>3,13,16</sup> although there is some discrepancy about the physical grounds of this approach.<sup>18,36</sup> LC- $\omega$ PBE and B3LYP yield very similar mean absolute errors (MAEs), although individual magnetic exchange couplings are somewhat different. Using the open-shell database of Valero and co-workers,<sup>13</sup> Rivero et al. have shown that the HSE functional is able to provide better magnetic exchange couplings when the SP instead of the NP approximation is used.<sup>37</sup> The reference database employed in that work consisted of 10 systems with two spin 1/2 magnetic centers: the H-He-H model system, two first-row compounds, and seven Cu-Cu complexes. Our results, however, do not show such conclusive evidence: Magnetic exchange couplings calculated with PBEh, HSE, and LC- $\omega$ PBE are comparable to those calculated with B3LYP, being the latter slightly better. However, if complex 8 is excluded, the B3LYP MAE in Table 1 is the largest of all the functionals included in this work. Notably, in this case, PBEh and HSE provide the best agreement with the experimental data, although all four functionals give close MAEs.

It should be pointed out that it is not the purpose of this work to assess the performance of different functionals against experimental data. Instead, we aim to compare the effect of range-separation on the calculation of magnetic exchange couplings. The comparison of our results with existing assessments suggest that larger test sets need to be employed to assess the performance of different methods for magnetic exchange couplings.

An implicit assumption in the evaluation of magnetic exchange couplings by equating the DFT and Heisenberg energy differences is that the DFT model is able to mimic the behavior of the Heisenberg model (eq 1), which implies

**Table 2.** Deviation from the Heisenberg Model As Given by the Parameter  $\eta(\times 10^{-3})$  as Defined in Equation 4<sup>a</sup>

complex	B3LYP	PBEh	HSE	LC- $\omega$ PBE
(1) Cu <sup>II</sup> -Cu <sup>II</sup>	1.67	2.69	2.75	2.97
(2) Cu <sup>II</sup> -Cu <sup>II</sup>	10.10	7.24	7.90	3.77
(3) Mn <sup>II</sup> -Cu <sup>II</sup>	9.52	7.41	7.81	5.96
(4) V <sup>IV</sup> -V <sup>IV</sup>	4.23	2.93	3.19	2.21
(5) Mn <sup>III</sup> Mn <sup>IV</sup>	22.05	16.85	17.35	16.02
(6) Cu <sup>II</sup> -Cu <sup>II</sup>	3.96	6.39	6.07	11.09
(7) Cu <sup>II</sup> -Cu <sup>II</sup>	4.73	4.56	4.56	4.15
(8) Cu <sup>II</sup> Cr <sup>III</sup>	26.56	28.16	84.90	10.61
(9) Cu <sup>II</sup> Mn <sup>III</sup>	40.98	14.15	18.65	0.29

<sup>a</sup> Smaller values of  $\eta$  indicate lesser variation of the local magnetization at the metallic centers between both high-spin and low-spin solutions.

that the magnetization at each center is the same for both spin configurations. Although this concept is difficult to quantify, a measure of the deviation of the DFT system from an ideal Heisenberg model can be given by the parameter

$$\eta = \left| \frac{S_A^{\text{HS}} S_B^{\text{HS}} + S_A^{\text{LS}} S_B^{\text{LS}}}{2(S_A^{\text{HS}} S_B^{\text{HS}} - S_A^{\text{LS}} S_B^{\text{LS}})} \right| \quad (4)$$

where  $S_{A,B}^{\text{HS,LS}}$  are the local magnetic moments (or integrated spin densities) at magnetic centers *A* and *B* for the HS and LS configurations. The parameter  $\eta$  is zero for the ideal case where both the HS and LS spin states hold the same local magnetization at both, *A* and *B*, magnetic centers. It should be commented that in the approximation proposed by Rudra et al.<sup>9,10</sup> the parameter  $\eta$  is exactly zero because of the constraint imposed in the local magnetic moments. Although there are several methods for partitioning the density and spin density into atomic contributions, we have chosen Mulliken population to estimate  $S_{A,B}^{\text{HS,LS}}$  since it is the most widely used method, although other partitioning methods based on local projectors<sup>38-41</sup> might be more suitable for large systems with many magnetic centers. Importantly, since the parameter  $\eta$  is based on differences of atomic spin populations between the LS and HS states, one would not expect that the values of  $\eta$  obtained using other population methods will follow the same trend for different density functionals. In Table 2 we show calculated values of  $\eta$  for all the complexes and density functionals employed in this



work. In all cases,  $S_{A,B}^{HS,LS}$  include atomic spin densities of the metallic centers and surrounding atoms with non-negligible magnetization. Overall, the parameter  $\eta$  is smaller for LC- $\omega$ PBE followed by PBEh, HSE, and B3LYP, in that order. This indicates that among the four hybrid functionals utilized in this work, B3LYP does the worst job in mimicking the Heisenberg behavior of all nine bimetallic complexes. This is in line with the observation of Rivero et al. that LC- $\omega$ PBE and HSE enhance the localization of the spin-density with respect to B3LYP, improving the description of spin localization (and hence magnetic exchange couplings) in these type of systems.<sup>37,42</sup>

## Conclusions

We have investigated the effect of Hartree–Fock range-separation on the calculation of magnetic exchange couplings by comparing magnetic exchange couplings using two global hybrid functionals, B3LYP and PBEh, with the short-range separated HSE and the long-range corrected LC- $\omega$ PBE in a test set of nine bimetallic transition-metal complexes containing 3d elements. Although our results show that there is no clear superiority of any of these functionals when comparing with experimental data, the LC- $\omega$ PBE provides a better description of the magnetization on the metallic centers, yielding self-consistent solutions for the high-spin and low-spin states that mimic more closely a Heisenberg-like behavior. The comparison of our results with existing assessments involving these same functionals separately suggest that larger test sets including all these functionals need to be employed to assess their performance for the prediction of magnetic exchange couplings.

**Acknowledgment.** JEP acknowledges support from NSF DMR-0906617 and an award from Research Corporation. J.I.M is member of CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina) and acknowledges financial support from Grants CONICET (PIP 5119/05). The authors thank T. van Voorhis and I. Rudra for kindly providing the crystallographic structures.

**Supporting Information Available:** Molecular data (spin configurations, total charge and spin multiplicities), atomic Cartesian coordinates, and complete basis sets. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## References

- Christou, G.; Gatteschi, D.; Hendrickson, D. N.; Sessoli, R. Single-molecule magnets. *MRS Bull.* **2000**, *25*, 66–71.
- Noodleman, L. Valence bond description of anti-ferromagnetic coupling in transition-metal dimers. *J. Chem. Phys.* **1981**, *74*, 5737–5743.
- Ruiz, E.; Alemany, P.; Alvarez, S.; Cano, J. Toward the prediction of magnetic coupling in molecular systems: Hydroxo- and alkoxo-bridged Cu(II) binuclear complexes. *J. Am. Chem. Soc.* **1997**, *119*, 1297–1303.
- Kortus, J.; Pederson, M. R.; Baruah, T.; Bernstein, N.; Hellberg, C. S. Density functional studies of single molecule magnets. *Polyhedron* **2003**, *22*, 1871–1876.
- Ruiz, E.; Cano, J.; Alvarez, S.; Alemany, P. Broken symmetry approach to calculation of exchange coupling constants for homobinuclear and heterobinuclear transition metal complexes. *J. Comput. Chem.* **1999**, *20*, 1391–1400.
- Noodleman, L.; Norman, J. G.; Osborne, J. H.; Aizman, A.; Case, D. A. Models for ferredoxins—Electronic-structures of iron sulfur clusters with one, two, and four iron atoms. *J. Am. Chem. Soc.* **1985**, *107*, 3418–3426.
- Ruiz, E.; Rodriguez-Forteza, A.; Cano, J.; Alvarez, S.; Alemany, P. About the calculation of exchange coupling constants in polynuclear transition metal complexes. *J. Comput. Chem.* **2003**, *24*, 982–989.
- Nishino, M.; Yamanaka, S.; Yoshioka, Y.; Yamaguchi, K. Theoretical approaches to direct exchange couplings between divalent chromium ions in naked dimers, tetramers, and clusters. *J. Phys. Chem. A* **1997**, *101*, 705–712.
- Rudra, I.; Wu, Q.; Van Voorhis, T. Accurate magnetic exchange couplings in transition-metal complexes from constrained density-functional theory. *J. Chem. Phys.* **2006**, *124*, 024103.
- Rudra, I.; Wu, Q.; Van Voorhis, T. Predicting exchange coupling constants in frustrated molecular magnets using density functional theory. *Inorg. Chem.* **2007**, *46*, 10539–10548.
- Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *J. Phys. Chem.* **1994**, *98*, 11623.
- Valero, R.; Costa, R.; Moreira, I.; Truhlar, D. G.; Illas, F. Performance of the M06 family of exchange-correlation functionals for predicting magnetic coupling in organic and inorganic molecules. *J. Chem. Phys.* **2008**, *128*, 114103.
- Polo, V.; Kraka, E.; Cremer, D. Electron correlation and the self-interaction error of density functional theory. *Mol. Phys.* **2002**, *100*, 1771–1790.
- Polo, V.; Kraka, E.; Cremer, D. Some thoughts about the stability and reliability of commonly used exchange-correlation functionals—Coverage of dynamic and nondynamic correlation effects. *Theor. Chem. Acc.* **2002**, *107*, 291–303.
- Ruiz, E.; Alvarez, S.; Cano, J.; Polo, V. About the calculation of exchange coupling constants using density-functional theory: The role of the self-interaction error. *J. Chem. Phys.* **2005**, *123*, 164110.
- Ruiz, E.; Cano, J.; Alvarez, S.; Polo, V. Reply to “Comment on ‘About the calculation of exchange coupling constants using density-functional theory: The role of the self-interaction error’” [*J. Chem. Phys.* **2005**, *123*, 164110]. *J. Chem. Phys.* **2006**, *124*, 107102.
- Adamo, C.; Barone, V.; Bencini, A.; Broer, R.; Filatov, M.; Harrison, N. M.; Illas, F.; Malrieu, J. P.; Moreira, I. D. R. Comment on “About the calculation of exchange coupling constants using density-functional theory: The role of the self-interaction error” [*J. Chem. Phys.* **2005**, *123*, 164110]. *J. Chem. Phys.* **2006**, *124*, 107101.
- Akande, A.; Sanvito, S. Exchange parameters from approximate self-interaction correction scheme. *J. Chem. Phys.* **2007**, *127*, 034112.
- Caballol, R.; Castell, O.; Illas, F.; Moreira, I.; Malrieu, J. Remarks on the proper use of the broken symmetry approach

- to magnetic coupling. *J. Phys. Chem. A* **1997**, *101*, 7860–7866.
- (21) Illas, F.; Moreira, I.; de Graaf, C.; Barone, V. Magnetic coupling in biradicals, binuclear complexes, and wide-gap insulators: A survey of ab initio wave function and density functional theory approaches. *Theor. Chem. Acc.* **2000**, *104*, 265–272.
- (22) Chevreau, H.; Moreira, I.; Silvi, B.; Illas, F. Charge density analysis of triplet and broken symmetry states relevant to magnetic coupling in systems with localized spin moments. *J. Phys. Chem. A* **2001**, *105*, 3570–3577.
- (23) Dai, D.; Whangbo, M. Spin exchange interactions of a spin dimer: Analysis of broken symmetry spin states in terms of the eigenstates of Heisenberg and Ising spin Hamiltonians. *J. Chem. Phys.* **2003**, *118*, 29–39.
- (24) Moreira, I. D. R.; Costa, R.; Filatov, M.; Illas, F. Restricted ensemble-referenced Kohn-Sham versus broken symmetry approaches in density functional theory: Magnetic coupling in Cu binuclear complexes. *J. Chem. Theory Comput.* **2007**, *3*, 764–774.
- (25) Heyd, J.; Scuseria, G. E.; Ernzerhof, M. Hybrid functionals based on a screened Coulomb potential. *J. Chem. Phys.* **2003**, *118*, 8207–8215.
- (26) Heyd, J.; Scuseria, G. E.; Ernzerhof, M. Hybrid functionals based on a screened Coulomb potential. *J. Chem. Phys.* **2006**, *124*, 219906.
- (27) Heyd, J.; Scuseria, G. E. Assessment and validation of a screened Coulomb hybrid density functional. *J. Chem. Phys.* **2004**, *120*, 7274–7280.
- (28) Krukau, A. V.; Vydrov, O. A.; Izmaylov, A. F.; Scuseria, G. E. Influence of the exchange screening parameter on the performance of screened hybrid functionals. *J. Chem. Phys.* **2006**, *125*, 224106.
- (29) Vydrov, O. A.; Scuseria, G. E.; Perdew, J. P. Tests of functionals for systems with fractional electron number. *J. Chem. Phys.* **2007**, *126*, 154109.
- (30) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Scalmani, G.; Mennucci, B.; Barone, V.; Petersson, G. A.; Caricato, M.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Li, X.; Hratchian, H. P.; Peralta, J. E.; Izmaylov, A. F.; Kudin, K. N.; Heyd, J. J.; Brothers, E.; Staroverov, V.; Zheng, G.; Kobayashi, R.; Normand, J.; Sonnenberg, J. L.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Burant, J. C.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Chen, W.; Wong, M. W.; Pople, J. A. *Gaussian Development Version*, revision F.02; Gaussian, Inc.: Wallingford CT, 2006.
- (31) Schafer, A.; Horn, H.; Ahlrichs, R. Fully optimized contracted Gaussian-basis sets for atoms Li to Kr. *J. Chem. Phys.* **1992**, *97*, 2571–2577.
- (32) Schafer, A.; Huber, C.; Ahlrichs, R. Fully optimized contracted Gaussian-basis sets of triple- $\zeta$  valence quality for atoms Li to Kr. *J. Chem. Phys.* **1994**, *100*, 5829–5835.
- (33) Perdew, J. P.; Ernzerhof, M.; Burke, K. Rationale for mixing exact exchange with density functional approximations. *J. Chem. Phys.* **1997**, *105*, 9982–9985.
- (34) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110*, 6158–6170.
- (35) Adamo, C.; Scuseria, G. E.; Barone, V. Accurate excitation energies from time-dependent density functional theory: Assessing the PBE0 model. *J. Chem. Phys.* **1999**, *111*, 2889–2899.
- (36) Illas, F.; Moreira, I. D. R.; Bofill, J. M.; Filatov, M. Extent and limitations of density functional theory in describing magnetic systems. *Phys. Rev. B* **2004**, *70*, 132414.
- (37) Rivero, P.; Moreira, I. D. R.; Illas, F.; Scuseria, G. E. Reliability of range-separated hybrid functionals for describing magnetic coupling in molecular systems. *J. Chem. Phys.* **2008**, *129*, 184110.
- (38) Clark, A. E.; Davidson, E. R. Local spin. *J. Chem. Phys.* **2001**, *115*, 7382–7392.
- (39) Clark, A.; Davidson, E. Population analyses that utilize projection operators. *Int. J. Quantum Chem.* **2003**, *93*, 384–394.
- (40) Mayer, I.; Hamza, A. Atomic decomposition of identity. General formalism for population analysis and energy decomposition. *Int. J. Quantum Chem.* **2005**, *103*, 798–807.
- (41) Mayer, I. Local spins. An alternative treatment for single determinant wave functions. *Chem. Phys. Lett.* **2007**, *440*, 357–359.
- (42) Rivero, P.; Moreira, I. D. R.; Scuseria, G. E.; Illas, F. Description of magnetic interactions in strongly correlated solids via range-separated hybrid functionals. *Phys. Rev. B* **2009**, *79*, 245129.
- (43) Song, H. H.; Zheng, L. M.; Liu, Y. J.; Xin, X. Q.; Jacobson, A. J.; Decurtins, S. Syntheses, structures, and magnetic properties of two copper(II) diphosphonates:  $[\text{NH}_3(\text{CH}_2)_2\text{NH}_3](2) [\text{Cu}-2(\text{hedp})(2)] \cdot \text{H}_2\text{O}$  and  $[\text{NH}_3\text{CH}(\text{CH}_3)\text{CH}_2\text{NH}_3](2) [\text{Cu}-2(\text{hedp})(2)]$  (hedp = 1-hydroxyethylidenediphosphonate). *J. Chem. Soc., Dalton Trans.* **2001**, *22*, 3274–3278.
- (44) Felthouse, T. R.; Laskowski, E. J.; Hendrickson, D. N. Magnetic exchange interactions in transition-metal dimers. 10. Structural and magnetic characterization of oxalate-bridged complex  $[\text{Cu}_2(\text{Et}_5\text{dien})_2(\text{C}_2\text{O}_4)](\text{Bph}_4)_2$  and related copper(II) dimers—Effects of nonbridging ligands and counterions on exchange interactions. *Inorg. Chem.* **1977**, *16*, 1077–1089.
- (45) Mathoniere, C.; Kahn, O.; Daran, J. C.; Hilbig, H.; Kohler, F. H. Complementarity and internal consistency between magnetic and optical-properties for the MnIICuII heterodinuclear compound  $[\text{Mn}(\text{Me}_6-[14]\text{Ane}-\text{N}_4)\text{Cu}(\text{Oxpn})](\text{CF}_3\text{SO}_3)_2$  ( $\text{Me}_6-[14]\text{Ane}-\text{N}_4 = (+/-)-5,7,7,12,14,14\text{-hexamethyl-1,4,8,11-tetraazacyclotetradecane}$ ; oxpn =  $n,n'$ -bis(3-aminopropyl)oxamide). *Inorg. Chem.* **1993**, *32*, 4057–4062.
- (46) Sun, Y.; Melchior, M.; Summers, D.; Thompson, R.; Rettig, S.; Orvig, C.  $[(\mu\text{-OCH}_3)\text{VO}(\text{ma})](2)$ , a strongly antiferromagnetic oxovanadium(IV) dimer. *Inorg. Chem.* **1998**, *37*, 3119–3121.
- (47) Sinnecker, S.; Neese, F.; Noodleman, L.; Lubitz, W. Calculating the electron paramagnetic resonance parameters of exchange coupled transition metal complexes using broken symmetry density functional theory: Application to a Mn-III/Mn-IV model compound. *J. Am. Chem. Soc.* **2004**, *126*, 2613–2622.

- (48) Tandon, S. S.; Thompson, L. K.; Manuel, M. E.; Bridson, J. N. Magnetostructural correlations in  $\mu(2)$ -1,1-*n*-3-bridged, dinuclear copper(II) complexes. 1. Ferromagnetic and antiferromagnetic coupling associated with the azide bridge—X-ray crystal-structures of [Cu-2(dmptd)( $\mu(2)$ -N-3)( $\mu(2)$ -cl)Cl-2]·CH<sub>3</sub>CN, [Cu-2(dmptd)( $\mu(2)$ -N-3)(2)(N-3)(2)], [Cu-2(dip)( $\mu(2)$ -N<sub>3</sub>)( $\mu(2)$ -Cl)Cl-2]·0.5CH<sub>3</sub>OH, [Cu-2(pap4Me-H)( $\mu(2)$ -N-3)(N-3)(2)]·0.33H<sub>2</sub>O, [Cu-2(pap)( $\mu(2)$ -N-3)Cl-3]·CH<sub>2</sub>Cl<sub>2</sub>, [Cu-2(pap)( $\mu(2)$ -N-3)(N-3)(NO<sub>3</sub>)(CH<sub>3</sub>OH)](NO<sub>3</sub>)·CH<sub>3</sub>OH, [Cu-2(ppd3me)( $\mu(2)$ -N-3)Cl-3(H<sub>2</sub>O)(1.5)], and [Cu-2(ppd)( $\mu(2)$ -N-3)(NO<sub>3</sub>)(3)(H<sub>2</sub>O)(1.6)]. *Inorg. Chem.* **1994**, *33*, 5555–5570.
- (49) Demunno, G.; Julve, M.; Lloret, F.; Faus, J.; Verdager, M.; Caneschi, A. Alternating ferromagnetic and antiferromagnetic interactions in unusual copper(II) chains. *Inorg. Chem.* **1995**, *34*, 157–165.
- (50) Birkelbach, F.; Winter, M.; Florke, U.; Haupt, H. J.; Butzlaff, C.; Lengen, M.; Bill, E.; Trautwein, A. X.; Wieghardt, K.; Chaudhuri, P. Exchange coupling in homodinuclear heterodinuclear complexes CuII<sub>2</sub> [M = Cr(III), Mn(III), Mn(II), Fe(III), Co(III), Co(II), Ni(II), Cu(II), Zn(II)]—Synthesis, structures, and spectroscopic properties. *Inorg. Chem.* **1994**, *33*, 3990–4001.

CT100104V

## Evaluations of the Absolute and Relative Free Energies for Antidepressant Binding to the Amino Acid Membrane Transporter LeuT with Free Energy Simulations

Chunfeng Zhao,<sup>†</sup> David A. Caplan,<sup>‡</sup> and Sergei Yu. Noskov<sup>\*,†</sup>

*Institute for Biocomplexity and Informatics and Department of Biological Sciences, University of Calgary, 2500 University Drive, B1558, Calgary, AB, Canada T2N 1N4 and Molecular Structure and Function, Hospital for Sick Children and Department of Biochemistry, University of Toronto, Ontario, Canada*

Received December 8, 2009

**Abstract:** The binding of ligands to protein receptors with high affinity and specificity is central to many cellular processes. The quest for the development of computational models capable of accurately evaluating binding affinity remains one of the main goals of modern computational biophysics. In this work, free energy perturbation/molecular dynamics simulations were used to evaluate absolute and relative binding affinity for three different antidepressants to a sodium-dependent membrane transporter, LeuT, a bacterial homologue of human serotonin and dopamine transporters. Dysfunction of these membrane transporters in mammals has been implicated in multiple diseases of the nervous system, including bipolar disorder and depression. Furthermore, these proteins are key targets for antidepressants including fluoxetine (aka Prozac) and tricyclic antidepressants known to block transport activity. In addition to being clinically relevant, this system, where multiple crystal structures are readily available, represents an ideal testing ground for methods used to study the molecular mechanisms of ligand binding to membrane proteins. We discuss possible pitfalls and different levels of approximation required to evaluate binding affinity, such as the dependence of the computed affinities on the strength of constraints and the sensitivity of the computed affinities to the particular partial charges derived from restrained electrostatic potential fitting of quantum mechanics electrostatic potential. Finally, we compare the effects of different constraint schemes on the absolute and relative binding affinities obtained from free energy simulations.

### I. Introduction

In last five years we have seen rapid and amazing progress in structural studies of membrane transporters. Several crystal structures for sodium-coupled membrane transporters have been solved at high resolution with and without ligand bound.<sup>1–4</sup> One of the first complexes of a membrane transporter with a bound drug was obtained for the LeuT–antidepressant complex. The structure, solved in a fully occluded state, contains bound ions, the transported solute,

and one of three tricyclic antidepressants (TCAs). LeuT is a bacterial leucine transporter, belonging to the large family of neurotransmitter sodium symporters (NSS).<sup>1</sup> Transporters of this family are involved in the termination of synaptic transmission through the reuptake of neurotransmitters (including glycine, glutamate, serotonin, dopamine, and many others) from the synapse into the cytoplasm of neurons and glia. Dysfunction of these membrane transporters in mammals has been implicated in multiple diseases of the nervous system.<sup>5</sup> Depression, one of the most prevalent psychiatric disorders, is directly associated with perturbation of serotonergic neurotransmission.<sup>6</sup> Antidepressants including fluoxetine (Prozac) and TCAs are known to bind membrane

\* Corresponding author. Telephone: (403) 210 7971. Fax (403) 220 8655. E-mail: snoskov@ucalgary.ca.

<sup>†</sup> University of Calgary.

<sup>‡</sup> University of Toronto.



transporters and block transport activity. Thus, understanding the mechanism of drug binding to these membrane transporters could possibly help the development of new therapeutics for depression. Recently, the crystal structures of LeuT bound to a variety of TCAs (clomipramine, CMI; imipramine, IMI; and desipramine, DSI) have been solved by two groups.<sup>2,3</sup> The TCAs bind in an extracellular-facing vestibule about 11 Å above the bound leucine substrate and the two sodium ions, as shown in the crystal structures.<sup>2</sup> It was demonstrated that they uncompetitively inhibit the transport of the leucine substrate, probably through the stabilization of the extracellular gate in a closed conformation.<sup>2,3</sup> With this in mind, the half-maximal inhibitory concentration (IC<sub>50</sub>) of these TCAs should strongly correlate to their binding affinities. In dose–response experiments reported by Singh et al.,<sup>2</sup> CMI has an IC<sub>50</sub> of inhibition of leucine transport of about eight-fold lower than IMI. Thermodynamically, this roughly corresponds to a free energy decrease of ~2 kT (about 1.5 kcal/mol at a temperature of 315 K). It was also shown that DSI is a less potent inhibitor compared to IMI (personal communication from S. Singh). Thus, Singh et al.'s data suggested the following affinity sequence: CMI > IMI > DSI. While ranking of the ligands for their binding affinity is available, it is difficult to measure absolute binding free energies for TCAs for their high nonspecific binding to chromatographic filters during separation, making evaluation of ratios between bound/unbound forms ambiguous. The closest measure to a  $K_d$  for TCA binding is the substrate uptake inhibition constant ( $K_i$ ) measured by radiolabeled uptake experiments, which depends on drug binding efficacy. These values alone do not represent an absolute  $K_d$  but will provide an accurate measure for relative potency of these drugs known for their ability to inhibit substrate transport and may be used to rank them accordingly. A recent development from the Javitch group on the use of a scintillation proximity assay (SPA) for measuring the  $K_d$  of radiolabeled compounds to detergent-solubilized material may lead to experimental  $K_d$  values in the future.<sup>7–9</sup> This technique was recently used to measure [<sup>3</sup>H]-citalopram binding to the presynaptic neuronal membrane serotonin transporter (SERT, homologous to LeuT), suggesting low  $\mu$ M to nM range of affinities.<sup>3,4</sup> The availability of experimental binding affinities,<sup>2</sup> together with the high-resolution crystal structures, make the TCA/LeuT system a rich platform for the testing and validation of various computational strategies for calculating binding free energies of drug binding.

The equilibrium thermodynamics of protein–ligand association is commonly described by binding affinity or Gibbs free energy ( $\Delta G$ ) and can be measured by a variety of standard biophysical or biochemical techniques, such as biospectroscopy, isothermal titration and differential scanning calorimetric techniques, electrophysiology, etc. However, to further our understanding of the process and its molecular determinants, it is important to obtain the quantitative contribution of different forces governing high affinity and specificity. Accurate prediction of binding affinity may, therefore, facilitate drug and protein design and optimization practices to attain better drugs with well-controlled binding

specificity and/or affinity. Therefore, the calculations of binding free energy by means of molecular simulations has been a major area of research in theoretical and computational chemistry/biochemistry<sup>10–20</sup> over the past 40 years. Many different approaches to evaluations of ligand affinities have been developed and can loosely be categorized into three major classes.<sup>21–24</sup> The first class of methods encompasses empirically driven schemes based on training sets derived from complexes with known structures.<sup>25,26</sup> The features of the known protein binding pocket can be translated into set of potential parameters used for virtual computer screening of large compound libraries or for designing novel ligands de novo.<sup>27,28</sup> Although it is a very powerful approach, its usability is limited if the system under study is lacking an extensive training set data.<sup>24</sup>

The second class of methods includes different extensions of popular molecular mechanics/Poisson–Boltzmann (generalized Born) surface area [MM/PB(GB)SA] algorithms,<sup>29,30</sup> where sampling of ligand/receptor coordinates achieved by molecular dynamics (MD) simulations and binding affinity is computed from collected trajectories.<sup>16,31–34</sup> The interaction energies (MM) are represented by respective force-field components for electrostatic and Lennard-Jones intermolecular terms, the nonelectrostatic component of the desolvation free energy is introduced via an empirical term (proportional to buried solvent accessible area), and the term accounting for the electrostatic penalty of water removal from the protein–ligand interface (desolvation penalty) is computed by means of continuum electrostatic models (such as Poisson–Boltzmann or generalized Born model). The collection of frames containing a protein–ligand complex can be extracted from MD/MC simulations, and each contribution can be averaged to obtain the binding free energies. This very popular and attractive method provides a straightforward and robust way for the computation of enthalpic contributions for a collection of frames extracted from MD but meets increasing difficulties when providing an accurate estimate for the loss in degrees of freedom incurred upon moving from bulk solution to the receptor-bound state and when accounting for dynamics/contribution due to explicit water present at the binding site.<sup>35,36</sup>

The third class includes approaches based on all-atom atomistic simulations.<sup>15</sup> The approaches vary from the application of thermodynamic integration<sup>37</sup> and free energy perturbation techniques to the computations of the potential of mean forces<sup>38</sup> with umbrella sampling methods,<sup>39</sup> adaptive biasing force MD,<sup>40</sup> or steered MD methods.<sup>41</sup> The difficulties associated with the requirement of sampling vast conformational space often lead to use of nonequilibrium simulations Jarzynski's equation,<sup>42</sup> modifications of Hamiltonian, such as metadynamics,<sup>43</sup> and a variety of enhanced sampling techniques.<sup>44</sup> Many of these methods require certain knowledge of the pathway of the drug binding/releasing. A very promising theoretical approach to the problem is to compute absolute binding free energies using a molecular dynamics/free energy perturbation (MD/FEP) method with constraining forces.<sup>35,45</sup> In this method, the free energies are calculated from the thermodynamic reversible work along an unphysical transformation path with MD/FEP

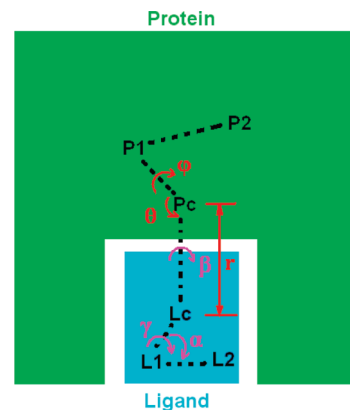
using a potential function that depends on one or several coupling parameters, such that the appropriate potential energy is recovered at the end-points.<sup>15,19,45–47</sup> Arguably, this approach provides the best way to describe the thermodynamics of protein–ligand recognition providing achievement of an adequate conformational sampling. The most common use of the FEP techniques is for the evaluation of the relative binding free energies between two different ligands through the perturbation of one ligand into another with a dual-topology scheme.<sup>17,48</sup> However, calculation of the absolute (standard) free energy arguably provides more detailed information about the mechanisms of ligand binding. It can also lead to the direct connection between macroscopic experimental measurements of binding affinities and the microscopic structural data extracted from FEP simulations.

To overcome expensive conformational sampling in computations of the absolute binding free energies with FEP/MD, one can use restraining potentials.<sup>18,35,45</sup> Recent applications of this technique have consisted of FEP/MD simulations with a reduced generalized solvent boundary condition (GSBP)<sup>49</sup> model with enhanced sampling by using conformational, translational, and orientational restraining potentials.<sup>35</sup> The method was developed and tested on a simple system of FKBP12/ligand<sup>35</sup> as well as a model system of T4 lysozyme/ligand<sup>45</sup> with considerable success. It has also been successfully applied to explain the substrate specificity of the LeuT neurotransmitter transporter,<sup>50</sup> emphasizing its applicability to membrane proteins as well. Despite rigorous theoretical foundation for FEP/MD methods and apparent success in applications studying highly specific binding to proteins,<sup>35</sup> the evaluation of many methodological aspects of these computations and its applicability to studies of ligand binding to membrane proteins is still a future goal.

In this report, we extend FEP/MD simulations to studies of an antidepressant binding to the membrane transporter LeuT. We are exploring the dependence of computed absolute free energies on the choice of atomic models, the strength of restraining potentials, and the reference structure for conformational constraints. We also report on the role of ligand reorganization free energy in high-affinity binding to the protein. It was recently suggested that contribution due to ligand reorganization upon transfer from bulk solution to a protein binding pocket could be as high as 7–8 kcal/mol and could potentially be the major determinant of high- vs low-affinity binding.<sup>28</sup> To test the validity of different approximations (PBC vs reduced GSBP approximation), we compare the free energy results from absolute free energy calculations to those from more common relative free energy calculations.<sup>17,48</sup> This report is organized as follows: Theoretical formulation of the FEP/MD absolute free energy calculation method is briefly reviewed in Section II. Computational details and results are presented in Sections III and IV, respectively. The conclusion of this report is summarized in Section V.

## II. Theoretical Formulation

**A. Restraint Forces.** To focus on relevant degrees of freedom, translational and orientational restraint potentials



**Figure 1.** Translational and rotational restraints on the ligand. Three positions in the protein (center of mass Pc, and randomly picked P1 and P2) and three positions in the ligand (center of mass Lc, and randomly picked L1 and L2) were used to set up the translational and rotational restraints. The translational restraints, shown in red, are defined by distance  $r$  (Pc and Lc), angle  $\theta$  (P1, Pc, and Lc), and dihedral  $\varphi$  (P2, P1, Pc, and Lc). The rotational restraints, shown in magenta, are defined by angle  $\alpha$  (L1, Lc, and Pc) and dihedrals  $\beta$  (P1, Pc, Lc, and L1) and  $\gamma$  (Pc, Lc, L1, and L2).

were implemented in all simulations, as described previously.<sup>18,35,45,51–53</sup> These two restraints are defined by three positions in the protein (Figure 1, center of mass Pc, and two randomly picked positions P1 and P2) and three positions in the ligand (center of mass Lc, and two randomly picked positions L1 and L2). The translational restraint is implemented to constrain the position of the center of mass of the ligand (Lc) relative to the protein. Its form is

$$u_t(r, \theta, \phi) = \frac{1}{2}k_{\text{dist}}(r - r_0)^2 + \frac{1}{2}k_{\text{ang}}(\theta - \theta_0)^2 + \frac{1}{2}k_{\text{ang}}(\phi - \phi_0)^2 \quad (1)$$

where  $r$  is the distance between Lc and Pc,  $\theta$  is the angle P1–Pc–Lc, and  $\varphi$  is the dihedral angle P2–P1–Pc–Lc. The corresponding reference values derived from an average of the equilibration trajectory are  $r_0$ ,  $\theta_0$ , and  $\varphi_0$ . The force constants for restraints on distance and angles (including dihedral angles) are  $k_{\text{dist}}$  and  $k_{\text{ang}}$ . Similarly, the rotational restraint on the ligand has the form of

$$u_r(\alpha, \beta, \gamma) = \frac{1}{2}k_{\text{ang}}(\alpha - \alpha_0)^2 + \frac{1}{2}k_{\text{ang}}(\beta - \beta_0)^2 + \frac{1}{2}k_{\text{ang}}(\gamma - \gamma_0)^2 \quad (2)$$

where  $\alpha$  is the angle Lc–L1–L2,  $\beta$  is the dihedral angle P1–Pc–Lc–L1, and  $\gamma$  is the dihedral angle Pc–Lc–L1–L2. The corresponding reference values derived from an average of the equilibration trajectory are  $\alpha_0$ ,  $\beta_0$ , and  $\gamma_0$ . The translational and rotational restraints ensure that the ligand is around its bound state. A configurational restraint ( $u_c$ ), in the form of a harmonic potential with respect to the root-mean-square deviation (RMSD) of the ligand, relative to a reference configuration, is also applied to constrain the ligand configuration.

**B. Standard Free Energy.** The free energy ( $\Delta G_b$ ) of a ligand (L) binding to a receptor protein (R) correlates to the equilibrium constant  $K_b$  of the binding reaction  $L + R \rightleftharpoons L \cdot R$  by

$$K_b = \frac{[L \cdot R]}{[L][R]} = \exp[-\beta \Delta G_b] \quad (3)$$

where  $\beta \equiv 1/(k_B T)$ , with  $k_B$  being the Boltzmann constant and  $T$  being the absolute temperature. Assuming the binding of a ligand is defined as moving one ligand molecule from the bulk solution to the binding site,  $K_b$  can be expressed as the following equation at low ligand concentration:

$$K_b = \frac{\int_{\text{site}} d\mathbf{L} \int d\mathbf{X} e^{-\beta U}}{\int_{\text{bulk}} d\mathbf{L} \delta(r_L - r^*) \int d\mathbf{X} e^{-\beta U}} \quad (4)$$

where  $\mathbf{L}$  and  $\mathbf{X}$  are the coordinates of the ligand molecule and the remaining atoms (including solvent, receptor protein, counterions, and others), respectively.  $U$  is the total potential energy of the system,  $r_L$  is the position of the center of mass of ligand L, and  $r^*$  is some arbitrary position in the bulk. The  $\delta$  function is a result of the translational invariance of the ligand in the bulk. To evaluate the binding free energy with molecular simulations, Deng et al.<sup>53</sup> wrote eq 4 in the form of the multiple of a series of intermediate states connecting the initial (the ligand in the binding site) and final (the ligand in the bulk) states:

$$K_b = \frac{\int_{\text{site}} d\mathbf{L} \int d\mathbf{X} e^{-\beta U_1}}{\int_{\text{site}} d\mathbf{L} \int d\mathbf{X} e^{-\beta(U_1+u_c)}} \times \frac{\int_{\text{site}} d\mathbf{L} \int d\mathbf{X} e^{-\beta(U_1+u_c)}}{\int_{\text{site}} d\mathbf{L} \int d\mathbf{X} e^{-\beta(U_1+u_c+u_t)}} \times \frac{\int_{\text{site}} d\mathbf{L} \int d\mathbf{X} e^{-\beta(U_1+u_c+u_t)}}{\int_{\text{site}} d\mathbf{L} \int d\mathbf{X} e^{-\beta(U_1+u_c+u_t+u_r)}} \times \frac{\int_{\text{site}} d\mathbf{L} \int d\mathbf{X} e^{-\beta(U_1+u_c+u_t+u_r)}}{\int_{\text{site}} d\mathbf{L} \int d\mathbf{X} e^{-\beta(U_0+u_c+u_t+u_r)}} \times \frac{\int_{\text{site}} d\mathbf{L} \int d\mathbf{X} e^{-\beta(U_0+u_c+u_t+u_r)}}{\int_{\text{bulk}} d\mathbf{L} \int d\mathbf{X} e^{-\beta(U_0+u_c+u_t)}} \times \frac{\int_{\text{bulk}} d\mathbf{L} \delta(r_L - r^*) \int d\mathbf{X} e^{-\beta(U_0+u_c)}}{\int_{\text{bulk}} d\mathbf{L} \delta(r_L - r^*) \int d\mathbf{X} e^{-\beta(U_0+u_c)}} \times \frac{\int_{\text{bulk}} d\mathbf{L} \delta(r_L - r^*) \int d\mathbf{X} e^{-\beta(U_0+u_c)}}{\int_{\text{bulk}} d\mathbf{L} \delta(r_L - r^*) \int d\mathbf{X} e^{-\beta(U_1+u_c)}} \times \frac{\int_{\text{bulk}} d\mathbf{L} \delta(r_L - r^*) \int d\mathbf{X} e^{-\beta(U_1+u_c)}}{\int_{\text{bulk}} d\mathbf{L} \delta(r_L - r^*) \int d\mathbf{X} e^{-\beta(U_1)}} \quad (5)$$

where the subscript 1 and 0 of  $U$  indicate fully interacting and fully decoupled ligand.

In terms of free energy contributions, the binding constant can be written as

$$K_b = \exp(+\beta \Delta G_c^{\text{site}}) \times \exp(+\beta \Delta G_r^{\text{site}}) \times \exp(+\beta \Delta G_t^{\text{site}}) \times \exp(-\beta \Delta G_{\text{int}}^{\text{site}}) \times F_r \times F_t \times \exp(+\beta \Delta G_{\text{int}}^{\text{bulk}}) \times \exp(-\beta \Delta G_c^{\text{bulk}}) \quad (6)$$

where the terms sequentially correspond to the terms in eq 5. All the terms involving  $\Delta G$  can be calculated by the standard free energy perturbation method,<sup>15,54,55</sup> while the free energy components associated with the configurational constraint ( $\Delta G_c^{\text{site}}$  and  $\Delta G_c^{\text{bulk}}$ ) can be better obtained by an umbrella sampling scheme and the translational ( $F_t$ ), and rotational ( $F_r$ ) factors can be evaluated directly with numerical integration schemes, since the interaction between the ligand molecule and the environment is turned off.

The standard binding free energy, defined relative to the standard concentration of 1 mol/L, is

$$\Delta G_b^0 \equiv -k_B T \ln[K_b C^0] = \Delta \Delta G_{\text{int}} + \Delta \Delta G_c + \Delta \Delta G_t^0 + \Delta \Delta G_r \quad (7)$$

where  $C^0 = 1$  mol/L and  $K_b C^0$  gives the standard binding constant. The free energy contributions are grouped as:  $\Delta \Delta G_{\text{int}} = \Delta G_{\text{int}}^{\text{site}} - \Delta G_{\text{int}}^{\text{bulk}}$ ,  $\Delta \Delta G_c = \Delta G_c^{\text{bulk}} - \Delta G_c^{\text{site}}$ ,  $\Delta \Delta G_t^0 = -\Delta G_t^{\text{site}} - k_B T \ln(F_t C^0)$ , and  $\Delta \Delta G_r = -\Delta G_r^{\text{site}} - k_B T \ln(F_r)$ , where  $\Delta \Delta G_{\text{int}}$  corresponds to the free energy difference associated with removing the ligand, restrained by the potential  $u_c$  from the bulk and inserting it to the binding site, restrained by  $u_c$ ,  $u_t$ , and  $u_r$ ;  $\Delta \Delta G_t^0$  and  $\Delta \Delta G_r$  correspond to the free energy changes associated with turning on and off the translational and rotational restraints on the ligand;  $\Delta \Delta G_c$  corresponds to free energy associated with the application of RMSD restraints. Recently, Deng and Roux further incorporated a grand canonical Monte Carlo step into the absolute binding free energy calculation to account for the fluctuation of the number of water molecules in highly occluded binding sites during alchemical perturbation.<sup>19,56</sup> For the TCA/LeuT systems, however, the TCA binding sites are open to access by water molecules from the extracellular side of the membrane. Thus, standard molecular dynamics trajectories offer adequate sampling for fluctuations in the number of water molecules around the binding site (Figure S2 in the Supporting Information).

**C. Decomposition of the Interaction Free Energy.** The interaction free energy,  $\Delta G_{\text{int}}^{\text{site}}$  and  $\Delta G_{\text{int}}^{\text{bulk}}$ , are decomposed into the contributions from electrostatic and Lennard-Jones (LJ) components. Further more, with the application of the Weeks–Chandler–Andersen (WCA) scheme,<sup>57</sup> the LJ potential, written as eq 8 in the CHARMM 27 all-atom force field, is uniquely separated into the repulsive (eq 9) and the dispersive (eq 10) potentials.

$$U^{\text{LJ}}(r) = \varepsilon \left[ \left( \frac{R_{\text{min}}}{r} \right)^{12} - 2 \left( \frac{R_{\text{min}}}{r} \right)^6 \right] \quad (8)$$

$$U^{\text{repu}}(r) = U^{\text{LJ}}(r) + \varepsilon \quad \text{when } r < R_{\text{min}};$$

$$U^{\text{repu}}(r) = 0 \quad \text{when } r \geq R_{\text{min}} \quad (9)$$

$$U^{\text{disp}}(r) = -\varepsilon \quad \text{when } r < R_{\text{min}}; \quad U^{\text{disp}}(r) = U^{\text{LJ}}(r)$$

$$\text{when } r \geq R_{\text{min}} \quad (10)$$

where  $\varepsilon$  has the dimension of energy, and  $R_{\text{min}}$  has the dimension of length. When the separation  $r$  of two atoms is at  $R_{\text{min}}$ , the LJ potential reaches its well depth  $-\varepsilon$ .

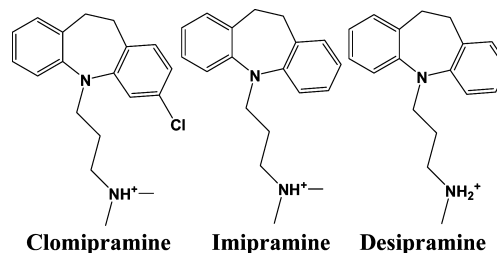
The interaction free energies ( $\Delta G_{\text{int}}^{\text{a}}$  where  $a$  represents *site* or *bulk*) are thus further separated into three components due to the contributions from electrostatic ( $\Delta G_{\text{elec}}^{\text{a}}$ ), dispersive LJ ( $\Delta G_{\text{disp}}^{\text{a}}$ ), and repulsive LJ ( $\Delta G_{\text{repu}}^{\text{a}}$ ) potentials.<sup>53</sup> This decomposition has shown to increase the statistical accuracy of the calculation of hydration free energies of molecules.<sup>53</sup> More importantly, while the decomposition of the interaction free energies is arbitrary to some degree, the values of these contributions helps to understand the nature of ligand binding.<sup>35,45,53,56</sup>

**D. Free Energy Perturbation.** The interaction free energies are evaluated by alchemical transformations using the standard free energy perturbation approach.<sup>15,54,55</sup> Briefly, the free energy contribution is calculated by gradually turning the potential on or off using a coupling parameter  $\lambda$  valued from 0 to 1. For example, to calculate the dispersive free energy for the ligand binding to the receptor binding site ( $\Delta G_{\text{disp}}^{\text{site}}$ ), a coupling parameter  $\lambda_{\text{disp}}$  is introduced. When  $\lambda_{\text{disp}} = 0$ , the dispersive interaction between the ligand and the environment is completely turned off, and when  $\lambda_{\text{disp}} = 1$ , the dispersive interaction between the ligand and the environment is completely turned on. The resulting auxiliary potential energy with the coupling constant is as follows

$$U(\lambda_{\text{disp}}) = U^0 + U^{\text{disp}}(\lambda_{\text{disp}}) \quad (11)$$

where  $U^0$  is the total potential when the dispersive interactions between the ligand and the environment are completely turned off, and  $U^{\text{disp}}(\lambda_{\text{disp}})$  is the total dispersive potential between the ligand and the environment scaled by the coupling constant. Several windows are applied to gradually increase the coupling constant from 0 to 1. For each window (from  $\lambda_{\text{disp},i}$  to  $\lambda_{\text{disp},j}$ ), the ensemble average  $\langle \exp \{-\beta[U(\lambda_{\text{disp},j}) - U(\lambda_{\text{disp},i})]\} \rangle_{U(\lambda_{\text{disp},i})}$  is obtained, and the sum of these windows corresponds to  $\exp(-\beta\Delta G_{\text{disp}}^{\text{site}})$ . Similar FEP/MD methods can be applied to calculate the other  $\Delta G$  components. For a detailed description of these FEP/MD procedures, readers are referred to the work by Deng et al.<sup>45</sup>

**E. Calculation of the Different Free Energy Components.** Combining the sequential process in eq 5 and the decomposition of the interaction energy described in section C, the steps corresponding to the dissociation of the fully interacting ligand in the protein binding site [system  $U_1(\text{site})$ ] as the initial state are listed in Table S1 in the Supporting Information. For each step, the second column gives the initial system, and the third column gives the system change. The corresponding free energy component and the method used to compute it are listed in the third and fourth columns, respectively. The free energies associated with the configurational constraints of the ligand to the reference configu-



**Figure 2.** Structural formulas of the three tricyclic antidepressants (TCAs) used for studies. The side-chain nitrogens are set in protonated form in accordance to the experimental conditions.

ration,  $\Delta G_{\text{c}}^{\text{site}}$  and  $\Delta G_{\text{c}}^{\text{bulk}}$ , are calculated by integrating the Boltzmann factor of the RMSD potential of mean force (PMF) obtained from umbrella sampling simulations. The translational factor ( $F_{\text{t}}$ ) and the rotational factor ( $F_{\text{r}}$ ) are numerically integrated from the expressions of

$$F_{\text{t}} = \int_0^{\infty} dr r^2 \int_0^{\pi} d\theta \sin(\theta) \int_{-\pi}^{\pi} d\phi \exp[-\beta u_{\text{t}}(r, \theta, \phi)] \quad (12)$$

$$F_{\text{r}} = \frac{1}{8\pi^2} \int_0^{\pi} d\alpha \sin(\alpha) \int_0^{\pi} d\beta \int_{-\pi}^{\pi} d\gamma \exp[-\beta u_{\text{r}}(\alpha, \beta, \gamma)] \quad (13)$$

where  $r, \theta, \varphi, \alpha, \beta, \gamma$  are the constrained internal coordinates illustrated in Figure 1, and  $u_{\text{t}}$  and  $u_{\text{r}}$  are the translational and rotational restraining potentials applied to the bound ligand presented in eqs 1 and 2. FEP/MD simulations are applied to get all of the other components of the absolute binding free energy.<sup>45</sup>

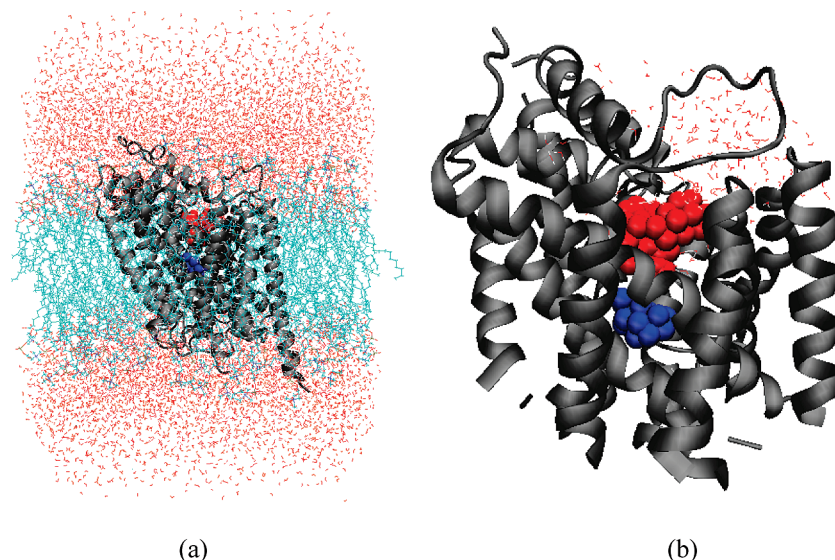
From Table S1 in the Supporting Information, it should be noted that when  $\Delta G_{\text{elec}}^{\text{site}}$ ,  $\Delta G_{\text{disp}}^{\text{site}}$ , and  $\Delta G_{\text{repu}}^{\text{site}}$  are calculated, restraints are applied during the alchemical transformations, thus making the values conditional on the restraining forces. Nonetheless, a comparison to their corresponding values in the bulk is informative about the binding.

### III. Methods

**A. Molecular Models of TCAs.** The chemical structures of the TCAs are shown in Figure 2. A neutral model for clomipramine (CMI) was originally developed in ref 58. New models with a positive charge for CMI, imipramine (IMI), and desipramine (DSI) were developed with the same procedure as described for CMI. Briefly, the geometric parameters (bond lengths, angles) were extracted from the crystallographic data. The positions of hydrogen atoms were unavailable from the crystal structures and were obtained by the HBUILD utility implemented with the CHARMM program.<sup>59</sup> The CHARMM27 force field<sup>60</sup> was used for the intramolecular potentials, and the nonbonded Lennard-Jones potential was used for the three drugs. Some of the torsional potentials are not available in the CHARMM27 force field and were obtained by fitting the B3LYP/6-31G\* torsion profiles.

Electrostatic potentials are crucial for reproducing the binding affinities of drugs to protein receptors. We assigned specific partial charges to each atom of the drugs using the





**Figure 3.** Simulation system for the TCA and substrate-bound leucine transporter LeuT. (a) The full simulation system for the MD equilibration. The protein is shown in cartoon mode (gray). A leucine substrate (blue) and a clomipramine drug (red) bound to the protein are shown in space-fill mode. The DPPC membrane (cyan) and water (red) molecules are shown in line mode, while the bound sodium ions and counterions (100 mM NaCl) are not shown. The whole system contains about 60 000 atoms. (b) A sphere containing LeuT (gray cartoon) and substrate (blue space-fill), the bound TCA (red space-fill), and water molecules (red lines) in the GSBP simulation system used for FEP/MD. Only atoms in a sphere (20 Å radius) centered on the ligand are represented explicitly. All atoms outside the sphere are represented implicitly using a continuum electrostatic approach.

RESP fitting approach described by Anisimov et al.<sup>61</sup> Briefly, for each drug, an electrostatic potential (ESP) grid was created on several Connolly surfaces<sup>62</sup> of the molecule by the CGRID program. ESP calculation at the B3LYP/6-31G\* level was applied to obtain the electrostatic map at the grid points, which was then used for partial charge fitting with the FITCHARGE module of the CHARMM program. The initial charge set ( $CS_{\text{initial}}$ ) was obtained for each atom type from the CHARMM nonpolarizable force field. A parabolic penalty function was used to restrain the values of the fitted charge set ( $CS_{\text{fitted}}$ ) with restraining forces of  $10^{-4} \text{ \AA}^2$  to the initial charge set ( $CS_{\text{initial}}$ ). Thus, the partial charges of  $CS_{\text{fitted}}$  have a better match to the electrostatic map. The initial ( $CS_{\text{initial}}$ ) and fitted ( $CS_{\text{fitted}}$ ) charges are presented in the Supporting Information. The total charge of each TCA molecule is +1 with the side-chain nitrogen protonated based on the reported  $pK_a$  values of the TCAs.<sup>63,64</sup>

**B. Equilibrium MD Simulation.** The starting configuration of drug and substrate-bound LeuT were taken from the X-ray coordinates revealed by Singh et al.<sup>2</sup> (Protein Data Bank entries 2Q6H, 2Q72, and 2QB4). The complexes were embedded in a lipid membrane using a multistep membrane-building procedure used in previous studies.<sup>50</sup> The simulation box contains the LeuT transporter, bound sodium and/or chloride ions, one leucine substrate, one antidepressant (clomipramine, imipramine, or desipramine) bound at the extracellular gate, and 148 dipalmitoylphosphatidylcholine (DPPC) lipid molecules solvated in an explicit 100 mM NaCl aqueous solution. A snapshot of the full simulation box is shown in Figure 3a. All computations were carried out by CHARMM, version c34b2, with the CHARMM27 force fields for proteins and lipids. MD simulation methods used here are similar to those used in previous studies of membrane systems.<sup>50</sup> Briefly, constant temperature/pressure

algorithms were applied (with pressure at 1 atm and temperature at 315 K). Periodic boundary conditions were used for the hexagonal system. Electrostatic interactions were treated with the particle mesh Ewald (PME) algorithm with a  $96 \times 96 \times 96 \text{ \AA}$  grid for fast Fourier transform,  $\kappa = 0.34 \text{ \AA}^{-1}$ , and a sixth-order spline interpolation. The nonbonded interactions were smoothly switched off at 12–14 Å. All simulation systems were equilibrated for 5 ns each without any configurational constraints.

**C. Absolute Binding Free Energy Calculation.** Following the equilibrium MD simulation, drug binding free energies were calculated using the protocol described in section II. To decrease computational cost, only the atoms in and surrounding the binding site (within 20 Å of the bound drug) were treated explicitly. All other atoms in the system were considered implicit, using a GSBP<sup>49</sup> generated for each system. It has been shown that the use of GSBP significantly decreases the size of the system (in our case from ~60 000 to ~7000 atoms, Figure 3b) while keeping the statistical error relatively low (~1–2 kcal/mol).<sup>35,45</sup> After the GSBP maps were generated, the reduced systems were minimized and equilibrated for 0.5 ns. Using the free energy decomposition protocol, the free energy components resulting from electrostatic, dispersive, repulsive, and constraining forces were measured independently. The CHARMM PERT function with the additional CHEMPERT option<sup>59</sup> was used for the FEP/MD simulations. All FEP/MD runs were equilibrated for 0.1 ns before collecting data during a 0.4 ns run. For the electrostatic component, both forward and reverse windows were calculated with the values for the coupling parameter  $\lambda$  set to [0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]. The dispersive component was measured through four forward windows with  $\lambda$  set to [0.0, 0.25, 0.75, 1.0]. The repulsive component was calculated by annihilating the

**Table 1.** Absolute Free Energy of Hydration for Clomipramine (C), Imipramine (I), and Desipramine (D)<sup>a</sup>

		$G_{\text{rep}}$	$G_{\text{disp}}$	$G_{\text{elec}}$	$G_{\text{conf}}$	$G_{\text{tot}}$	$G_{\text{conf\_vac}}$	$G_{\text{solv}}$
CS <sub>fitted</sub>	C	39.4 ± 0.5	-35.3 ± 0.2	-50.7 ± 0.1	-4.9 ± 0.6	<b>-51.4 ± 0.4</b>	-3.1	-48.3
	I	38.3 ± 0.5	-33.6 ± 0.2	-51.4 ± 0.2	-5.5 ± 0.8	<b>-52.2 ± 0.8</b>	-7.2	-45.1
	D	37.3 ± 0.9	-32.4 ± 0.2	-56.2 ± 0.1	-3.6 ± 1.2	<b>-54.9 ± 1.9</b>	-5.0	-49.9
CS <sub>initial</sub>	C	39.9 ± 0.8	-35.0 ± 0.2	-54.9 ± 0.1	-7.2 ± 1.4	<b>-57.2 ± 2.4</b>	-10.0	-47.1
	I	38.5 ± 0.7	-33.6 ± 0.2	-55.5 ± 0.3	-4.6 ± 0.8	<b>-55.2 ± 0.6</b>	-9.3	-45.8
	D	37.3 ± 0.8	-32.2 ± 0.1	-62.3 ± 0.1	-3.9 ± 0.1	<b>-61.1 ± 0.8</b>	-8.5	-52.6

<sup>a</sup> The data are obtained with the application of SSBP. The first group (rows 2–4) of results is obtained from simulations with the fitted charge set (CS<sub>fitted</sub>), and the second group (rows 5–7) of results is obtained with the initially guessed charge set (CS<sub>initial</sub>).  $G_{\text{tot}}$  is the sum of the free energy components of repulsive ( $G_{\text{rep}}$ ), dispersive ( $G_{\text{disp}}$ ), electrostatic ( $G_{\text{elec}}$ ), and conformational ( $G_{\text{conf}}$ ).

molecule from binding site or bulk solvent with the application of a soft-core potential.<sup>53</sup> The annihilation is done from both forward and reverse windows with  $\lambda$  set to [0.0, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]. The translational and rotational constraint components were measured through forward windows with  $\lambda$  set to [0.0, 0.0025, 0.0050, 0.0075, 0.01, 0.02, 0.04, 0.06, 0.08, 0.1, 0.2, 0.4, 0.6, 0.8, 1.0]. Similar coupling constants were applied in the work of Deng et al.<sup>45</sup> and Wang et al.<sup>35</sup> The results were processed with the weighted histogram analysis method<sup>65</sup> (the WHAM module of CHARMM) to remove any bias due to the restraining forces. Hydration free energies of the TCAs were calculated by FEP/MD, using a model system of TCA solvated by 400 water molecules. The spherical solvent boundary potential (SSBP)<sup>53,66</sup> was applied to account for the influence outside of the solvation sphere. The FEP protocol used to determine hydration free energy is the same as that described above for the computation of absolute binding free energy. Equilibration without constraints was performed for 100 ps, and window lengths for evaluation of free energy were 200 ps. Statistical uncertainty is reported with the standard deviations of each free energy component and the total solvation, site, and binding free energies obtained through separating the FEP/MD simulations to three blocks and obtaining the free energies for each block using WHAM analysis.<sup>65</sup>

**D. Relative Binding Free Energy Calculation.** The binding free energy difference between CMI and IMI as well as IMI and DSI were calculated with the CHARMM PERT function. For example, in the case of CMI/IMI, the equilibrated membrane system of LeuT/CMI (equilibrated from PDB entry 2Q6H) is used as the starting configuration ( $\lambda = 0$ ), and in the final configuration, CMI is perturbed to IMI ( $\lambda = 1$ ). For the perturbation, 11 windows were used varying between 0.0 and 1.0 by increments of 0.1. For each perturbation window, a 10 ps equilibration run and a 190 ps production run were applied. We also use the equilibrated membrane system of LeuT/IMI (equilibrated from PDB entry 2Q72), as the  $\lambda = 0$  state and similar procedures were carried out to calculate the relative binding free energy by perturbing IMI to CMI. The relative binding free energies from the forward and backward simulations are within 1 kcal/mol of difference, and the average relative free energy is reported.

## IV. Results and Discussion

**A. Hydration Free Energy of TCAs.** The hydration free energies of the three TCAs are listed in Table 1.  $G_{\text{tot}}$ , the sum of the free energy components of repulsive ( $G_{\text{rep}}$ ),

dispersive ( $G_{\text{disp}}$ ), electrostatic ( $G_{\text{elec}}$ ) and conformational ( $G_{\text{conf}}$ ), will be subtracted from  $G_{\text{tot}}(\text{site})$  (Table 2) to obtain the absolute binding free energy. To obtain the hydration free energy, the component caused by RMSD constraining force for the solute in vacuum ( $G_{\text{conf\_vac}}$ ) needs to be subtracted from  $G_{\text{tot}}$ . The resulting hydration free energies for all the three TCAs from the calculations with SSBP potential are in the order of  $-50$  kcal/mol. The calculations indicate that the electrostatic component ( $G_{\text{elec}}$ ), compared to the LJ component ( $G_{\text{repu}} + G_{\text{disp}}$ ) of the free energy, mostly accounts for the favorable hydration free energies for the three TCAs.

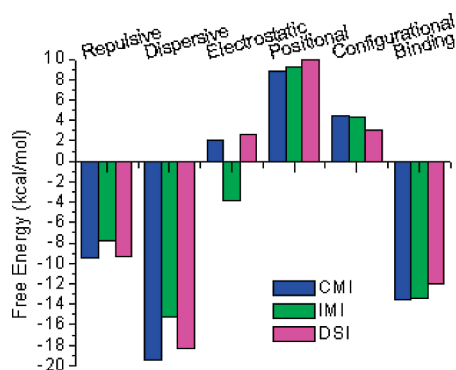
The influence of applied partial charge sets on the solvation free energy is apparent. While there are only slight differences (within 1 kcal/mol) on solvation free energies for the two charge sets (CS<sub>fitted</sub> and CS<sub>initial</sub>) of CMI and IMI, there is a decrease of about 3 kcal/mol in hydration free energy of DSI from CS<sub>fitted</sub> to CS<sub>initial</sub>. The most significant change is the electrostatic free energy component which has been decreased from  $-56.2$  for CS<sub>fitted</sub> to  $-62.3$  kcal/mol for CS<sub>initial</sub>. This indicates that, given relatively small changes in the partial charges (Table S2, columns 7 to 6 in the Supporting Information), the electrostatic contribution of hydration free energy could be dramatic. It should be noted, however, that the hydration free energies obtained here are for drug transfer from vacuum to bulk solution. The solubility data reported for many pharmaceutical agents is not readily comparable to theoretical results, as one has to first correct for the drug sublimation free energy.

**B. Absolute Binding Free Energies of TCAs Binding to LeuT.** The computed absolute binding free energies of TCAs to LeuT are listed in Table 2. The absolute free energy of binding is the difference between the free energies of site binding and hydration ( $G_{\text{tot}}$  in Table 1). As described, the absolute energy is decomposed into the contributions from electrostatic, dispersive, and repulsive parts as well as from constraint potentials on RMSD, orientation, and translation. The final absolute binding free energies are  $-13.6$ ,  $-13.4$ , and  $-12.0$  kcal/mol for CMI, IMI, and DSI, respectively. All of them are negative, indicating favorable binding. As discussed in Section II, while the decomposition of the free energies is path dependent, the comparison between the same contributions for different ligands can still provide valuable information about the nature of binding (Figure 4). CMI has a lower binding free energy than IMI mainly due to its gain in the dispersive free energy ( $-4.1$  kcal/mol), which may be due to favorable interactions achieved by its additional chlorine with the LeuT side chain.<sup>2</sup> Despite the overall

**Table 2.** Absolute Free Energy of Binding for Clomipramine (C), Imipramine (I), and Desipramine (D) to LeuT in Reduced GSBP System<sup>a</sup>

			$G_{\text{rep}}$	$G_{\text{disp}}$	$G_{\text{elec}}$	$G_{\text{pos}}$			$G_{\text{conf}}$	$G_{\text{tot}}$
						$G_{\text{const}}$	$-k_{\text{B}}T \ln(F_{\text{r}}C^0)$	$-k_{\text{B}}T \ln(F_{\text{r}})$		
CS <sub>fitted</sub>	site	C	30.0 ± 0.7	-54.6 ± 0.3	-48.7 ± 0.6	-3.0 ± 0.1	5.4	6.5	-0.5 ± 0.1	-64.9 ± 1.2
		I	30.5 ± 1.1	-48.8 ± 0.4	-55.3 ± 0.8	-2.6 ± 0.1	5.4	6.4	-1.2 ± 0.1	-65.6 ± 0.4
		D	28.0 ± 0.4	-50.8 ± 0.1	-53.6 ± 0.1	-2.1 ± 0.1	5.6	6.4	-0.5 ± 0.1	-67.0 ± 0.6
	binding	C	-9.4 ± 0.9	-19.4 ± 0.6	2.0 ± 0.6		8.8 ± 0.1		4.4 ± 0.5	-13.6 ± 1.4
		I	-7.8 ± 1.4	-15.3 ± 0.6	-3.9 ± 0.9		9.3 ± 0.1		4.3 ± 0.8	-13.4 ± 1.0
		D	-9.3 ± 0.4	-18.3 ± 0.1	2.6 ± 0.1		9.9 ± 0.1		3.0 ± 1.2	-12.1 ± 1.4
CS <sub>initial</sub>	site	C	30.6 ± 0.9	-54.4 ± 0.3	-54.3 ± 0.4	-1.5 ± 0.1	5.4	6.5	-0.5 ± 0.1	-68.2 ± 0.6
		I	33.1 ± 0.2	-48.3 ± 0.1	-58.3 ± 0.4	-1.9 ± 0.1	5.4	6.4	-1.6 ± 0.1	-65.2 ± 0.4
		D	26.4 ± 1.1	-49.0 ± 0.5	-60.2 ± 0.6	-2.6 ± 0.1	5.4	6.4	-1.3 ± 0.1	-74.9 ± 1.7
	binding	C	-9.3 ± 1.5	-19.4 ± 0.2	0.6 ± 0.3		10.3 ± 0.1		6.7 ± 1.3	-11.1 ± 2.5
		I	-5.5 ± 0.7	-14.7 ± 0.3	-2.7 ± 0.3		9.9 ± 0.1		3.0 ± 0.7	-10.0 ± 0.7
		D	-10.9 ± 1.6	-16.8 ± 0.4	2.1 ± 0.6		9.2 ± 0.2		2.6 ± 0.2	-13.8 ± 2.1

<sup>a</sup> The first group (rows 2–7) of results is obtained from simulations with the fitted charge set (CS<sub>fitted</sub>), and the second group (rows 8–13) of results is obtained with the initially guessed charge set (CS<sub>initial</sub>). For the site free energy,  $G_{\text{tot}}$  is the sum of the free energy components of repulsive ( $G_{\text{rep}}$ ), dispersive ( $G_{\text{disp}}$ ), electrostatic ( $G_{\text{elec}}$ ), positional ( $G_{\text{pos}}$ , including translational and rotational constraints), and conformational ( $G_{\text{conf}}$ ). For the binding free energy, the values are obtained by subtracting the corresponding components of solvation free energy (Table 1) from the site free energy. Note that CS<sub>initial</sub> is the initial charge set based on CHARMM force field, and the atomic charges were derived from similar atom types in the CHARMM27 force field and were not parametrized. CS<sub>fitted</sub> is the RESP charge fitted for the electrostatic density map from QM calculation, and it reflects the chemical environment of the atoms of the drugs better.



**Figure 4.** Column illustration of the binding free energy components of three TCA's (CMI in blue, IMI in green, and DSI in magenta) binding to LeuT. From left to right: repulsive, dispersive, electrostatic, positional (including translational and rotational constraints), configurational, and total binding free energy.

unfavorable binding free energy for DSI compared to IMI, the DSI binding free energy exhibits a more favorable repulsive component. This is explained by the fact that DSI has one less methyl group on its “tail” and is accommodated better by the binding pocket.

The complete expression for the binding constants (eq 4–5) contains a term that describes conformational dynamics of the receptor. To perform their function, membrane transporters may undergo large conformational changes, binding and unbinding ions, and opening and closing extracellular and intracellular gates. Such events take place over large time intervals in the  $\mu\text{s}$  to  $\text{s}$  range. These changes in the protein's structure are not present in the ns MD/free energy simulations. Only one state of the transporter (the “occluded” state) is considered, leading, therefore, to a large overestimation of the absolute drug binding affinities. The contribution of the binding site's conformational changes to the complete partition function is expected to be unfavorable. This problem is well-known, and the interested reader may refer to an excellent review by Mobley and Dill.<sup>67</sup> Clearly,

this component plays a dominant role in differences between computed and experimental binding affinities. However, assuming that these large conformational changes are ion induced and independent of the particular drug, one may conclude that relative free energies or drug ranking based on the absolute free energy computations will be robust, since the term describing receptor dynamics will cancel out. Regardless, the absolute binding free energies, though lacking in contribution from receptor allosteric changes, provide molecular insights on key factors governing formation of the high-affinity/-specificity complex between the protein and the drug. Below we will provide detailed discussion on the importance of different factors in computations of absolute binding affinities with FEP/MD Simulations.

**C. Restraints of Different Strengths.** The application of biasing restraints on configuration, translation, and orientation of the ligand greatly reduces the configuration space and enhances the sampling efficiency. However, the choice of the restraining force constants has been shown to affect the outcome of individual free energy contributions (i.e., dispersive, repulsive, electrostatic, etc.), despite the resulting binding free energy being largely unaffected.<sup>35</sup> In this section, several sets of different force constants are used in calculating the absolute free energy of CMI binding to LeuT. The effect of these contributions on the total absolute free energy is evaluated. From Table 3, the distance constant ( $k_{\text{r}}$ , in kcal/mol/Å<sup>2</sup>) has little effect on the value of  $G_{\text{int}}(\text{site})$ , the sum of electrostatic, dispersive, and repulsive free energies as well as the total absolute binding free energy, which is evident from the comparison of the results from ( $k_{\text{c}} = 10$ ,  $k_{\text{t}} = 10$ , and  $k_{\text{a}} = 200$ ) to ( $k_{\text{c}} = 10$ ,  $k_{\text{t}} = 1$ , and  $k_{\text{a}} = 200$ ). Reducing the strength of the angular and dihedral force constant ( $k_{\text{a}}$ , in kcal/mol/rad<sup>2</sup>) makes  $G_{\text{int}}(\text{site})$  more unfavorable but only slightly affects the final binding free energy. Lowering the RMSD force constant ( $k_{\text{c}}$ , in kcal/mol/Å<sup>2</sup>) to 1 kcal/mol/Å<sup>2</sup> only slightly changes both  $G_{\text{int}}$  and the final binding free energy, as evident in the set ( $k_{\text{c}} = 1$ ,  $k_{\text{t}} = 10$ , and  $k_{\text{a}} = 200$ ). Thus, the choice of the constraining forces is robust to some



**Table 3.** Computed Binding Free Energy for the Clomipramine/LeuT Complex at Different Force Constants for the RMSD Potentials, the Translational Restraint, and the Rotational Restraint<sup>a</sup>

$k_c:k_t:k_a$	$G_{\text{repu}}(\text{site})$	$G_{\text{disp}}(\text{site})$	$G_{\text{elec}}(\text{site})$	$-G_{\text{const}}(\text{site})$	$-k_B T \ln(F_t C^0)$	$-k_B T \ln(F_r)$	$-G_{\text{conf}}(\text{site})$	$G_{\text{tot}}^0(\text{site})$	$\Delta G_{\text{binding}}^0$
10:10:200	30.0 ± 0.7	-54.6 ± 0.3	-48.7 ± 0.6	-3.0 ± 0.1	5.4	6.5	-0.5 ± 0.1	-64.9 ± 1.2	-13.6 ± 1.4
1:1:20	32.5 ± 0.8	-52.2 ± 0.5	-48.4 ± 0.6	-0.3 ± 0.1	3.2	4.3	-0.1	-61.0 ± 0.7	-13.3 ± 1.0
10:1:200	30.9 ± 1.2	-54.2 ± 0.5	-48.3 ± 0.3	-1.4 ± 0.1	4.4	6.2	-0.4 ± 0.1	-62.8 ± 1.0	-12.0 ± 0.5
10:10:20	32.2 ± 3.4	-53.8 ± 0.3	-48.9 ± 0.2	-0.4 ± 0.1	3.9	4.3	-0.5	-63.1 ± 3.2	-10.3 ± 2.1
1:10:200	32.2 ± 2.6	-53.8 ± 0.4	-49.1 ± 0.7	-1.6 ± 0.1	5.1	6.2	-0.1	-61.0 ± 3.2	-13.8 ± 3.1
100:100:2000	28.6 ± 1.0	-55.0 ± 0.4	-49.1	-4.3 ± 0.1	7.5	8.6	-47.4 ± 0.9	-111.1 ± 1.0	-18.3 ± 1.0
0:0:0	38.9 ± 1.7	-53.0 ± 0.3	-47.6 ± 0.4	N/A	N/A	N/A	N/A	-61.7 ± 2.2	-15.4 ± 1.8

<sup>a</sup> The values  $k_c$  (in kcal/mol/Å<sup>2</sup>),  $k_t$  (in kcal/mol/Å<sup>2</sup>), and  $k_a$  (in kcal/mol/rad<sup>2</sup>) are the force constants for the RMSD potentials, the distance force constant for the translational restraint, and the angle/dihedral force constant for the translational and rotational restraints, respectively.

**Table 4.** Absolute Free Energy of Binding for Clomipramine (C), Imipramine (I), and Desipramine (D) to LeuT in Reduced GSBP System<sup>a</sup>

		$G_{\text{rep}}$	$G_{\text{disp}}$	$G_{\text{elec}}$	$G_{\text{pos}}$			$G_{\text{conf}}$	$G_{\text{tot}}$
					$G_{\text{const}}$	$-k_B T \ln(F_t C^0)$	$-k_B T \ln(F_r)$		
site	C	36.5 ± 0.6	-54.2 ± 0.2	-48.9 ± 0.8	-3.0 ± 0.2	5.4	6.5	-10.7 ± 0.3	-68.4 ± 0.8
	I	32.3 ± 1.3	-49.2 ± 0.1	-51.3 ± 0.5	-2.6 ± 0.1	5.4	6.4	-5.3 ± 0.3	-64.3 ± 1.6
	D	30.3 ± 0.7	-51.7 ± 0.2	-52.2 ± 0.7	-2.6 ± 0.1	5.6	6.4	-5.1 ± 0.4	-69.3 ± 0.9
solv	C	40.4 ± 0.3	-35.9 ± 0.3	-51.7 ± 0.1		N/A		-3.5 ± 1.0	-50.7 ± 1.6
	I	39.1 ± 0.6	-34.0 ± 0.2	-51.5 ± 0.1		N/A		-4.4 ± 0.5	50.9 ± 0.3
	D	36.7 ± 0.7	-32.6 ± 0.3	-56.0 ± 0.1		N/A		-2.7 ± 0.2	-54.6 ± 0.6
bind-ing	C	-3.9 ± 0.9	-18.3 ± 0.4	2.7 ± 0.8		8.8 ± 0.2		-7.2 ± 0.8	-17.9 ± 1.5
	I	-6.7 ± 0.8	-15.2 ± 0.2	0.3 ± 0.4		9.3 ± 0.1		-0.9 ± 0.8	-13.2 ± 1.3
	D	-6.4 ± 0.4	-19.0 ± 0.3	3.9 ± 0.7		9.4 ± 0.1		-2.4 ± 0.4	-14.5 ± 1.3

<sup>a</sup> The average favorable solvated structures of the drugs are used as the reference structures for the site and hydration free energy calculations. The numbers are reported in kcal/mol.

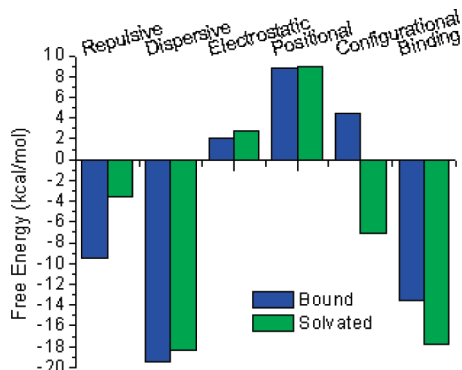
degree, as the calculations with several sets of constraining constants produce similar binding free energies around -13.6 kcal/mol. Nevertheless, we picked  $k_c = 10$  kcal/mol/Å<sup>2</sup>,  $k_t = 10$  kcal/mol/Å<sup>2</sup>, and  $k_a = 200$  kcal/mol/rad<sup>2</sup> as the set of constraining constants, as it generates relatively low statistical errors for each contribution of the free energies. This set of force constants is consistent with the one applied by Deng et al.<sup>45</sup> The combination of  $k_c = 10$  kcal/mol/Å<sup>2</sup>,  $k_t = 1$  kcal/mol/Å<sup>2</sup>, and  $k_a = 200$  kcal/mol/rad<sup>2</sup> provides lower binding free energy error, but the free energy component for repulsive interaction is increased which indicates that the low error in binding free energy might not be sustainable. Interestingly, the hardest ( $k_c = 100$ ,  $k_t = 100$ , and  $k_a = 2000$ ) and softest ( $k_c = 0$ ,  $k_t = 0$ , and  $k_a = 0$ ) constraints applied lead to the largest statistical deviations in binding free energy (-18.3 and -15.4 kcal/mol, respectively), reflecting two major problems one may face with constrained FEP simulations: under-sampled conformational space for substrate dynamics and over-restricted decreased conformational space that prohibits substrate dynamics in the site.

**D. Effect of the Reference Structure Choice and RMSD Constraint Scheme.** Up to this point we have used the average bound structure as the reference structure for the configurational constraint. The free energy component due to this constraint has a positive sign which indicates that there is a free energy loss upon binding due to the fact that the configurational freedom is restricted when one drug is bound to the receptor (moving from solvent to the protein binding pocket). A less natural, but still reasonable, reference structure is the lowest-energy structure for the hydrated drug. To find out the preferred conformation in the bulk solution for three drugs, we have performed an extensive replica exchange MD simulation.<sup>68,69</sup> Briefly, the trajectories of the

replica exchange MD simulation of solvated drugs were used to obtain a probability distribution of conformations according to RMSD. The solvated structures of drugs with the most probable RMSDs were averaged to give the referenced average-solvated structure. The free energies calculated using these reference structures are listed in Table 4. The binding free energy changes from -13.6 to -17.9 for CMI, from -13.4 to -13.2 for IMI, and from -12.0 to -14.5 kcal/mol for DSI. CMI is still stands out as the most potent inhibitor for LeuT. DSI becomes more favorable than IMI, but the difference is within the statistical uncertainty (2.6 kcal/mol), considering that IMI and DSI have very similar binding affinity. However, the free energy differences between CMI and IMI are more pronounced and are significantly higher than that of estimated IC50 values reported from the experiment.<sup>2</sup>

It is interesting to examine the variation of the free energy components due to the choice of reference structures for the drugs (Figure 5). The sum of nonbonding components of the relative free energy ( $G_{\text{rep}} + G_{\text{disp}} + G_{\text{elec}}$ ) becomes unfavorable when using the average solvated structure as the reference structure for conformational constraints. This is due to the fact that the reference structure is not the average bound structure, and the bound state differs substantially from that found in the bulk. The free energy component from the configurational constraint changes its sign to a negative value, compensating for the unfavorable constraint of restraining the ligand to the bulk-optimized conformation. Thus, by using the average solvated structure as a reference structure, we calculate the  $G_{\text{int}}(\text{site})$  with an unfavorable bound structure, and we must rely the  $G_{\text{conf}}(\text{site})$  component to correct the result. This is reflected by the standard mean deviations of the  $G_{\text{conf}}(\text{site})$ . When the average bound





**Figure 5.** Column illustration of the binding free energy components of CMI binding to LeuT calculated with the average bound structure (blue) and the average favorable solvated structure (green) as the configurational reference of the RMSD constraint. From left to right: repulsive, dispersive, electrostatic, positional (including translational and rotational constraints), configurational, and total binding free energy.

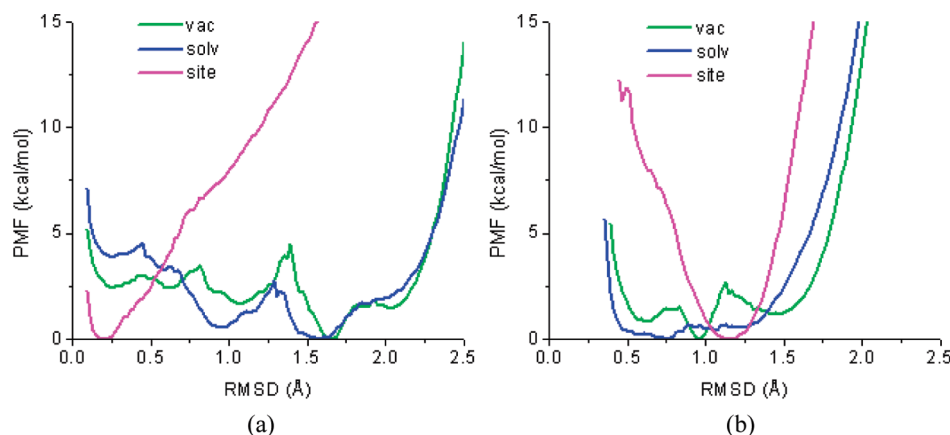
structure is used, the standard deviation of  $G_{\text{conf}}(\text{site})$  is 0.1 kcal/mol (Table 2), while when the favorable solvated structure is used, the standard deviation of  $G_{\text{conf}}(\text{site})$  is about 0.3–0.4 kcal/mol (Table 4). Contrary to this, by using the average bound structure as a reference structure, we calculate  $G_{\text{in}}(\text{solv})$  with an unfavorable solvated structure, and we must rely the  $G_{\text{conf}}(\text{solv})$  component to correct the result. Since we usually have a more restricted space for the ligand in the binding pocket than in the bulk,  $G_{\text{conf}}(\text{solv})$  should converge better than  $G_{\text{conf}}(\text{site})$  (i.e., more overlapping between windows of umbrella sampling). Thus, using the average bound structure as a reference structure would usually be a better choice despite the fact that one could also get reasonable results with the average solvated structure as a reference.

To illustrate this compensation due to the use of different reference structures, we have evaluated the potential of mean force (PMF) as a function of the RMSD constraint. The PMFs of the configurational restraints for clomipramine are shown in Figure 6, where the reference structure of the restraint is the average bound structure in (a) and the average most probable solvated structure in (b). When the average bound structure is the reference structure, the free energy component resulting from the configurational restraint (4.4 kcal/mol) can be understood by a comparison of the site and solvent PMFs. The PMF of clomipramine in solvent goes up to about 4.4 kcal/mol at a RMSD corresponding to the minimum of the PMF in the binding site ( $\sim 0.2$  Å). Similarly, when the average most probable solvated structure is the reference structure, the free energy component resulting from the configurational restraint ( $-7.2$  kcal/mol) can also be intuitively understood. The PMF of clomipramine in the binding site goes up to about 7 kcal/mol at a RMSD, corresponding to the minimum of the PMF in solvent ( $\sim 0.7$  Å). That is, in order to keep the configuration of bound structure around its most probable solvated structure, a free energy of 7 kcal/mol is required. Recently, Yang et al. incorporated the ligand reorganization free energy emphasizing its importance and even occasionally dominant contribution to high-affinity binding. It was shown that accounting

for the free energy change for ligands between the free and bound states leads to better binding affinity prediction and enhanced the correlation coefficient for a number of studied complexes.<sup>28</sup> While they obtained the ligand reorganization free energy in the frame of MM-GBSA (molecular mechanics-generalized Born surface area), we want to point out here that a similar free energy component due to the reorganization of the ligand can also be robustly introduced within the FEP/MD absolute binding free energy calculation scheme.

**E. Sensitivity to Potential Parameters.** There are typically no molecular models of drugs available for the study of drug/protein binding free energy. For all-atom molecular simulations, it is a common practice to get the equilibrium structure from the crystal structure, and by using quantum mechanical calculations, determine the atomic charges and/or intramolecular potentials and obtain the remaining parameters from an available force field. An important concern is the assignment of atomic charges. Deng et al.<sup>45</sup> showed that with CHARMM, CHELPG, and AMSOL,<sup>70</sup> the calculated binding free energies could differ by about 3 kcal/mol for aromatic molecules binding to the T4 Lysozyme L99A mutant. In this report, we examine the dependence of calculated binding free energies on two charge sets: one initially guessed set based on CHARMM force field ( $\text{CS}_{\text{initial}}$ ) and one charge set obtained from the RESP scheme described by Anisimov et al.<sup>61</sup> ( $\text{CS}_{\text{fitted}}$ ). The method of charge fitting for  $\text{CS}_{\text{fitted}}$  is described in Section III. The atomic charges for the two charge sets are listed in Table S2 in the Supporting Information. The influence of using different charge sets ( $\text{CS}_{\text{fitted}}$  and  $\text{CS}_{\text{initial}}$ ) was shown to be significant when calculating the hydration free energies of the TCAs (a difference of 6.2 kcal/mol for  $G_{\text{tot}}(\text{solv})$  of DSI). On the other hand, both sets produced similar binding free energies for the ligands. Nevertheless, Table 2 shows that, from  $\text{CS}_{\text{fitted}}$  to  $\text{CS}_{\text{initial}}$ , the total binding free energy (in kcal/mol) changed from  $-13.6$  to  $-11.0$  for CMI, from  $-13.4$  to  $-10.0$  for IMI, and from  $-12.0$  to  $-13.8$  for DSI. For this particular binding site and these ligands, the different charge sets exhibit a difference up to 3.4 kcal/mol. Notably, the free energy order also changes, now contradicting the experimental results. Thus, for ligands without partial charge parameters in the available force field, the derivation of partial charges from high-level QM electrostatic potentials is highly preferred for molecular models, and validation against available experimental data is warranted.

**F. Comparison between GSBP and PBC Simulations. Relative Binding Free Energies of TCAs Binding to LeuT.** To compare possible artifacts due to reduction of the system with the GSBP scheme, we performed atomistic free energy simulations for the full system embedded into lipid bilayer using the FEP technique. The relative binding free energies can be calculated between two pairs: CMI/IMI and IMI/DSI from FEP simulations or simply from differences in absolute binding free energies obtained with GSBP simulations. The relative binding free energies between these compounds are known, and thus it is possible to correlate performance of two methods (GSBP and PBC) to experimental data. The structural difference between CMI and IMI is that the chlorine atom in CMI is replaced by a hydrogen



**Figure 6.** PMF of the configurational restraints on clomipramine. The PMFs of clomipramine in the binding site (site), solvent (solv), and vacuum (vac) are shown in magenta, blue, and green, respectively. The reference structure of clomipramine is the average bound structure (a) and the average favorable solvating structure (b).



**Figure 7.** Free energy cycle of the binding of CMI, IMI, and DSI to LeuT. The binding free energies of CMI ( $\Delta G(\text{binding, CMI})$ ), IMI ( $\Delta G(\text{binding, IMI})$ ), and DSI ( $\Delta G(\text{binding, DSI})$ ) are obtained from absolute binding free energy calculations. The relative free energies ( $\Delta G(\text{solv})$  and  $\Delta G(\text{binding})$ ) are obtained from relative binding free energy calculations. The numbers are reported in kcal/mol.

atom in IMI. The difference between IMI and DSI is an addition of a methyl group on the tail (see Figure 2). The results are shown in Figure 7. The relative free energy for CMI to IMI is 6.0 kcal/mol for the bound state and 5.6 kcal/mol for the unbound state (in bulk), leading to a relative binding free energy of about 0.4 kcal/mol. The result is remarkably comparable to the difference of the absolute binding free energy of CMI to IMI, 0.2 kcal/mol obtained from absolute binding free energy simulations. The relative free energy for IMI to DSI is  $-16.5$  and  $-17.5$  kcal/mol for the bound and unbound states, respectively, leading to a relative binding free energy of 1.0 kcal/mol, which is also comparable to the difference of the absolute binding free energy of IMI to DSI of 1.4 kcal/mol. This, to some degree, justifies the use of GSBP for the absolute binding free energy calculations compared to the periodical boundary conditions used in the relative binding free energy calculations. The results are very interesting as one can use the more mature relative binding free energy calculation to check the results of absolute binding free energies when experimental binding affinities are not available. It also suggests that the reduced system provides an excellent and, more importantly, a relatively cheap test ground for rapid evaluation of relative free energies (for ranking of substrates). The thermodynamic cycle can be applied as an assessment tool to further the development of absolute binding free energy calculation methodologies.

### G. Implications for the Molecular Mechanism of Antidepressant Binding to LeuT.

Crystal structures of TCA-bound LeuT<sup>2,3</sup> show that the substrate and the drug binding sites are quite close to each other, mainly separated by a charged pair Arg 30 and Asp 404. Singh et al. demonstrated that the binding of the substrate and CMI might be thermodynamically coupled.<sup>2</sup> To explore this possibility, we calculated the absolute binding free energies of TCA's binding to LeuT without the bound leucine substrate. As crystal structures for such systems are not directly available, we obtained the starting structures by removing the substrates from the TCA-bound LeuT (2Q6H, 2Q72, 2QB4 after step B in section III) and allowing a further 2 ns relaxation and equilibration of the structures. The results are presented in Table 5. Generally, the binding free energies for all three drugs decreases when the substrates are removed: from  $-13.4$  to  $-16.9$  for CMI, from  $-13.2$  to  $-15.0$  for IMI, and from  $-12.0$  to  $-13.4$  kcal/mol for DSI. For the most bulky drug, CMI, the repulsive van der Waals (vdW) interaction is the main reason for the more favorable binding free energies. The free energy contribution due to repulsive vdW interactions changes from  $-9.4$  for the substrate-bound LeuT (Table 2) to  $-17.0$  kcal/mol for the substrate-free LeuT (Table 5). One explanation is that the removal of the substrate relaxes the protein structure, allowing for a more flexible drug binding pocket, facilitating the binding of CMI. The results here immediately infer that it is very difficult to compare the computed free energies to their corresponding experimental values, since experimental binding affinities reflect the binding free energies of a combination of LeuT with and without substrate.

Asp 401 is one of the key residues in the TCA binding pocket of LeuT.<sup>2,3</sup> The presence of the charge in this position is preserved in several NSS transporters. It is sometimes switched to a positively charged lysine or arginine and is thought to be functionally important. Its negatively charged side-chain carboxylate forms a salt bridge with the protonated side-chain nitrogen atom (Figure 2) in the "tail" of the bound TCAs.<sup>2,3</sup> To quantify the contribution of this interaction, we mutated the negatively charged Asp 401 to a positively charged lysine residue using the SCWRL3.0 program<sup>71</sup>

**Table 5.** Absolute Free Energy of Binding for Clomipramine (C), Imipramine (I), and Desipramine (D) to Substrate-Free LeuT in Reduced GSBP System

		$G_{\text{rep}}$	$G_{\text{disp}}$	$G_{\text{elec}}$	$G_{\text{pos}}$			$G_{\text{conf}}$	$G_{\text{tot}}$
					$G_{\text{const}}$	$-k_{\text{B}}T \ln(F_{\text{I}}C^0)$	$-k_{\text{B}}T \ln(F_{\text{I}})$		
site	C	23.1 ± 1.0	-51.7 ± 0.5	-49.4 ± 0.2	-1.7 ± 0.1	5.3	6.5	-0.5 ± 0.1	-68.5 ± 1.1
	I	31.6 ± 1.8	-48.5 ± 0.6	-57.2 ± 0.3	-2.2 ± 0.1	5.5	6.4	-1.6 ± 0.1	-66.0 ± 1.7
	D	25.9 ± 0.9	-47.8 ± 0.6	-53.8 ± 0.3	-2.1 ± 0.1	5.4	6.4	-1.9 ± 0.3	-67.9 ± 0.4
binding	C	-17.0 ± 1.6	-16.8 ± 0.7	1.0 ± 0.3		10.0 ± 0.1		5.9 ± 0.8	-16.9 ± 1.5
	I	-6.9 ± 2.2	-14.8 ± 0.7	-5.8 ± 0.3		9.7 ± 0.1		2.8 ± 0.7	-15.0 ± 2.1
	D	-10.9 ± 0.7	-15.7 ± 0.4	2.2 ± 0.2		9.7 ± 0.1		1.3 ± 0.5	-13.4 ± 0.8

starting from the equilibrated, membrane embedded 2Q6H structure. We thus obtained the D401K mutant of LeuT with bound ions, substrate, and clomipramine. The calculated absolute binding free energy for CMI binding to LeuT–D401K is  $-9.9$  kcal/mol with contributions of  $-17.2$ ,  $-19.7$ ,  $13.2$ ,  $10.2$ , and  $3.6$  kcal/mol from the repulsive, dispersive, electrostatic, translational, rotational, and configurational components, respectively. Comparing CMI binding between the mutant and wild-type systems, the loss in the electrostatic interaction is  $11.2$  kcal/mol, primarily due to lack of interaction between CMI's fully protonated side-chain nitrogen atom and the negative side-chain of D401 residue in the wild-type LeuT. This loss is only partially compensated by the gains of  $-7.8$  kcal/mol from the repulsive term. It may be due to the fact that the loss of the salt bridge allows CMI to rearrange in the binding pocket and avoid strong repulsive interactions. It should be noted that the overall effect of the mutation is a  $3.7$  kcal/mol less favorable for binding. This result has recently received surprising support from the combination of electrophysiological and biochemical studies on related GABA transporters. Cherubino et al. have reported that charge-switching mutations of lysine at the position 448 (K448E and K448D) that corresponds to 401 in LeuT lead to significant increase in the efficacy of desipramine,<sup>72</sup> further supporting functional role of negative charge in this location for high-affinity antidepressant binding.

## V. Conclusion

In this report we used the free energy perturbation/molecular dynamics (FEP/MD) method with constraints to calculate the standard binding free energies for tricyclic antidepressant (TCA) binding to LeuT. The computed binding free energies are comparable to the experimental results. We showed that restraining potentials are essential and robust for enhancing sampling in studies of drugs binding to membrane proteins. The choice of the magnitude of the restraining forces on translation, rotation, and configuration are relatively robust (within 2 kcal/mol), as long as extreme values are avoided. For the configurational constraint, we tried two kinds of reference structures: the average bound structure and the average favorable solvated structure. It was shown that the use of the bound structure as the reference structure produced better results. Thus it is recommended for future applications of the FEP/MD method for the evaluation of absolute binding free energies. The use of the generalized solvent boundary potential (GSBP) approximation for studies of drug binding to membrane proteins also appears to be justified and

accurate. We also showed that developing the molecular mechanics models (charge sets) for drugs using quantum mechanical electrostatic potential maps provided better accuracy. Interestingly, there is a notable compensation observed in the binding free energies between both molecular mechanics models. The absolute free energies (such as hydration and binding free energies) may vary significantly across models, while the relative binding free energies between drugs differ only by a small amount across models. This result is encouraging and in good agreement with similar conclusions from discussions on ion–protein interactions and force field development. We also showed that with the current FEP/MD method for absolute binding free energy, the results are compatible with those from the relative binding free energy calculations. We propose that the application of the free energy cycle can be applied to assess new methods of absolute binding free energy calculations. Finally, we showed that the absolute binding free energies for TCAs binding to LeuT are slightly different in the substrate-bound and substrate-free situations, indicating substrate-drug coupling as proposed by Singh et al.<sup>2,3</sup> Consistent with experimental indication, our FEP/MD calculation of the absolute binding free energy showed that the D401K mutation impairs the binding of clomipramine to LeuT, proving the essential role of the salt bridge between D401 of LeuT and the protonated nitrogen in the “tail” of the TCAs. In conclusion, this report shows that the use of the FEP/MD method for calculating absolute free energies of drugs bound to membrane proteins is a promising tool that can be used for drug design.

**Acknowledgment.** We are gratefully acknowledged discussions with Drs. Benoit Roux, Julia Subbotina, and Yuqing Deng. We are greatly indebted to Satinder Singh for guidance and discussion of the experimental work on LeuT and LeuT complexations with antidepressants. This work was supported by a Discovery Grant from the Natural Sciences and Engineering Council of Canada (NSERC) to S.N. S.N. is a CIHR New Investigator, an Alberta Heritage Foundation for Medical Research Scholar, and an Alberta Ingenuity New Scholar. The computational support for this work was provided by the West-Grid Canada through a resource allocation award to S.N. C.F.Z. is an AHFMR Post-Doctoral Fellow.

**Supporting Information Available:** Steps for the calculation of different components of the absolute free energy of a ligand binding to a receptor site. Molecular model for partial-charge assignment and atom labeling in clomi-



pramine. RESP ( $CS_{\text{fitted}}$ ) and initial ( $CS_{\text{initial}}$ ) partial charges for clomipramine, imipramine, and desipramine. Figure shows the binding pocket before and after the interactions between clomipramine (CMI) and the environment are fully turned off. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## References

- (1) Yamashita, A.; Singh, S. K.; Kawate, T.; Jin, Y.; Gouaux, E. Crystal structure of a bacterial homologue of Na<sup>+</sup>/Cl<sup>-</sup>-dependent neurotransmitter transporters. *Nature* **2005**, *437*, 215.
- (2) Singh, S. K.; Yamashita, A.; Gouaux, E. Antidepressant binding site in a bacterial homologue of neurotransmitter transporters. *Nature* **2007**, *448*, 952.
- (3) Zhou, Z.; Zhen, J.; Karpowich, N. K.; Goetz, R. M.; Law, C. J.; Reith, M. E. A.; Wang, D. N. LeuT-desipramine structure reveals how antidepressants block neurotransmitter reuptake. *Science* **2007**, *317*, 1390.
- (4) Zhou, Z.; Zhen, J.; Karpowich, N. K.; Law, C. J.; Reith, M. E. A.; Wang, D. N. Antidepressant specificity of serotonin transporter suggested by three LeuT-SSRI structures. *Nat. Struct. Mol. Biol.* **2009**, *16*, 652.
- (5) Rudnick, G. Serotonin transporters - Structure and function. *J. Membr. Biol.* **2006**, *213*, 101.
- (6) Schafer, W. R. How do antidepressants work? Prospects for genetic analysis of drug mechanisms. *Cell* **1999**, *98*, 551.
- (7) Quick, M.; Winther, A. M. L.; Shi, L.; Nissen, P.; Weinstein, H.; Javitch, J. A. Binding of an octylglucoside detergent molecule in the second substrate (S2) site of LeuT establishes an inhibitor-bound conformation. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 5563.
- (8) Shi, L.; Quick, M.; Zhao, Y. F.; Weinstein, H.; Javitch, J. A. The mechanism of a neurotransmitter: sodium symporter - Inward release of Na<sup>+</sup> and substrate is triggered by substrate in a second binding site. *Mol. Cell* **2008**, *30*, 667.
- (9) Quick, M.; Javitch, J. A. Monitoring the function of membrane transport proteins in detergent-solubilized form. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 3603.
- (10) Beveridge, D. L.; Dicapua, F. M. Free-Energy Via Molecular Simulation - Applications to Chemical and Biomolecular Systems. *Annu. Rev. Biophys. Biomol. Struct.* **1989**, *18*, 431.
- (11) Jorgensen, W. L. Free-Energy Calculations - a Breakthrough for Modeling Organic-Chemistry in Solution. *Acc. Chem. Res.* **1989**, *22*, 184.
- (12) Sneddon, S. F.; Tobias, D. J.; Brooks, C. L. Thermodynamics of Amide Hydrogen-Bond Formation in Polar and Apolar Solvents. *J. Mol. Biol.* **1989**, *209*, 817.
- (13) Jorgensen, W. L.; Severance, D. L. Aromatic Aromatic Interactions - Free-Energy Profiles for the Benzene Dimer in Water, Chloroform, and Liquid Benzene. *J. Am. Chem. Soc.* **1990**, *112*, 4768.
- (14) Tobias, D. J.; Brooks, C. L. The Thermodynamics of Solvophobic Effects - a Molecular-Dynamics Study of Normal-Butane in Carbon-Tetrachloride and Water. *J. Chem. Phys.* **1990**, *92*, 2582.
- (15) Kollman, P. Free-Energy Calculations - Applications to Chemical and Biochemical Phenomena. *Chem. Rev.* **1993**, *93*, 2395.
- (16) Gilson, M. K.; Given, J. A.; Bush, B. L.; McCammon, J. A. The statistical-thermodynamic basis for computation of binding affinities: A critical review. *Biophys. J.* **1997**, *72*, 1047.
- (17) Simonson, T.; Archontis, G.; Karplus, M. Free energy simulations come of age: Protein-ligand recognition. *Acc. Chem. Res.* **2002**, *35*, 430.
- (18) Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M. Absolute binding free energies: A quantitative approach for their calculation. *J. Phys. Chem. B* **2003**, *107*, 9535.
- (19) Deng, Y. Q.; Roux, B. Computations of Standard Binding Free Energies with Molecular Dynamics Simulations. *J. Phys. Chem. B* **2009**, *113*, 2234.
- (20) Zhou, H. X.; Gilson, M. K. Theory of Free Energy and Entropy in Noncovalent Binding. *Chem. Rev.* **2009**, *109*, 4092.
- (21) Guvench, O.; MacKerell, A. D. Computational evaluation of protein-small molecule binding. *Curr. Opin. Struct. Biol.* **2009**, *19*, 55.
- (22) Villoutreix, B. O.; Bastard, K.; Sperandio, O.; Fahraeus, R.; Poyet, J. L.; Calvo, F.; Deprez, B.; Miteva, M. A. In silico-in vitro screening of protein-protein interactions: Towards the next generation of therapeutics. *Curr. Pharm. Biotechnol.* **2008**, *9*, 103.
- (23) Gane, P. J.; Dean, P. M. Recent advances in structure-based rational drug design. *Curr. Opin. Struct. Biol.* **2000**, *10*, 401.
- (24) Gilson, M. K.; Zhou, H. X. Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21.
- (25) Gohlke, H.; Hendlich, M.; Klebe, G. In *Knowledge-based scoring function to predict protein-ligand interactions*. *J. Mol. Biol.* **2000**, *295*, 337.
- (26) Sondergaard, C. R.; Garrett, A. E.; Carstensen, T.; Pollastri, G.; Nielsen, J. E. Structural Artifacts in Protein-Ligand X-ray Structures: Implications for the Development of Docking Scoring Functions. *J. Med. Chem.* **2009**, *52*, 5673.
- (27) Klebe, G. Recent developments in structure-based drug design. *J. Mol. Med.* **2000**, *78*, 269.
- (28) Yang, C. Y.; Sun, H. Y.; Chen, J. Y.; Nikolovska-Coleska, Z.; Wang, S. M. Importance of Ligand Reorganization Free Energy in Protein-Ligand Binding-Affinity Prediction. *J. Am. Chem. Soc.* **2009**, *131*, 13709.
- (29) Wang, J. M.; Morin, P.; Wang, W.; Kollman, P. A. Use of MM-PBSA in reproducing the binding free energies to HIV-1 RT of TIBO derivatives and predicting the binding mode to HIV-1 RT of efavirenz by docking and MM-PBSA. *J. Am. Chem. Soc.* **2001**, *123*, 5221.
- (30) Bashford, D.; Case, D. A. Generalized born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129.
- (31) Swanson, J. M. J.; McCammon, J. A. Applying the statistical mechanics behind ligand-receptor binding affinities. *Biophys. J.* **2003**, *84*, 341A.
- (32) Minh, D. D. L.; Bui, J. M.; Chang, C. E.; Jain, T.; Swanson, J. M. J.; McCammon, J. A. The entropic cost of protein-protein association: A case study on acetylcholinesterase binding to fasciculin-2. *Biophys. J.* **2005**, *89*, L25.
- (33) Wright, J. D.; Noskov, S. Y.; Lim, C. Factors governing loss and rescue of DNA binding upon single and double mutations in the p53 core domain. *Nucleic Acids Res.* **2002**, *30*, 1563.



- (34) Noskov, S. Y.; Lim, C. Free energy decomposition of protein-protein interactions. *Biophys. J.* **2001**, *81*, 737.
- (35) Wang, J. Y.; Deng, Y. Q.; Roux, B. Absolute binding free energy calculations using molecular dynamics simulations with restraining potentials. *Biophys. J.* **2006**, *91*, 2798.
- (36) Irudayam, S. J.; Henschman, R. H. Entropic Cost of Protein-Ligand Binding and Its Dependence on the Entropy in Solution. *J. Phys. Chem. B* **2009**, *113*, 5871.
- (37) Kirkwood, J. G. Statistical mechanics of fluid mixtures. *J. Chem. Phys.* **1935**, *3*, 300.
- (38) Roux, B. The Calculation of the Potential of Mean Force Using Computer-Simulations. *Comput. Phys. Commun.* **1995**, *91*, 275.
- (39) Frenkel, D.; Smit, B. *Understanding molecular simulation: from algorithms to applications*, 2nd ed.; Academic Press: San Diego, CA, 2002; p xxii.
- (40) Darve, E.; Pohorille, A. Calculating free energies using average force. *J. Chem. Phys.* **2001**, *115*, 9169.
- (41) Gullingsrud, J. R.; Braun, R.; Schulten, K. Reconstructing potentials of mean force through time series analysis of steered molecular dynamics simulations. *J. Comput. Phys.* **1999**, *151*, 190.
- (42) Jarzynski, C. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.* **1997**, *78*, 2690.
- (43) Ensing, B.; De Vivo, M.; Liu, Z. W.; Moore, P.; Klein, M. L. Metadynamics as a tool for exploring free energy landscapes of chemical reactions. *Acc. Chem. Res.* **2006**, *39*, 73.
- (44) Christi, C. D.; Mark, A. E.; van Gunsteren, W. F. Basic ingredients of free energy calculations: A review. *J. Comput. Chem.* **2010**, *31*, 1569.
- (45) Deng, Y. Q.; Roux, B. Calculation of standard binding free energies: Aromatic molecules in the T4 lysozyme L99A mutant. *J. Chem. Theory Comput.* **2006**, *2*, 1255.
- (46) Brandsdal, B. O.; Osterberg, F.; Almlof, M.; Feierberg, I.; Luzhkov, V. B.; Aqvist, J. Free energy calculations and ligand binding. *Protein Simul.* **2003**, *66*, 123.
- (47) Nervall, M.; Hanspers, P.; Carlsson, J.; Boukharta, L.; Aqvist, J. Predicting binding modes from free energy calculations. *J. Med. Chem.* **2008**, *51*, 2657.
- (48) Archontis, G.; Watson, K. A.; Xie, Q.; Andreou, G.; Chrysina, E. D.; Zographos, S. E.; Oikonomakos, N. G.; Karplus, M. Glycogen phosphorylase inhibitors: A free energy perturbation analysis of glucopyranose spirohydantoin analogues. *Proteins: Struct., Funct., Bioinf.* **2005**, *61*, 984.
- (49) Im, W.; Berneche, S.; Roux, B. Generalized solvent boundary potential for computer simulations. *J. Chem. Phys.* **2001**, *114*, 2924.
- (50) Noskov, S. Y. Molecular mechanism of substrate specificity in the bacterial neutral amino acid transporter LeuT. *Proteins: Struct., Funct., Bioinf.* **2008**, *73*, 851.
- (51) Straatsma, T. P.; Zacharias, M.; Mccammon, J. A. Holonomic Constraint Contributions to Free-Energy Differences from Thermodynamic Integration Molecular-Dynamics Simulations. *Chem. Phys. Lett.* **1992**, *196*, 297.
- (52) Boresch, S. The role of bonded energy terms in free energy simulations - Insights from analytical results. *Mol. Simul.* **2002**, *28*, 13.
- (53) Deng, Y. Q.; Roux, B. Hydration of amino acid side chains: Nonpolar and electrostatic contributions calculated from staged molecular dynamics free energy simulations with explicit water molecules. *J. Phys. Chem. B* **2004**, *108*, 16567.
- (54) Zwanzig, R. W. High-Temperature Equation of State by a Perturbation Method 0.1. Nonpolar Gases. *J. Chem. Phys.* **1954**, *22*, 1420.
- (55) Straatsma, T. P.; Mccammon, J. A. Computational Alchemy. *Annu. Rev. Phys. Chem.* **1992**, *43*, 407.
- (56) Deng, Y. Q.; Roux, B. Computation of binding free energy with molecular dynamics and grand canonical Monte Carlo simulations. *J. Chem. Phys.* **2008**, 128.
- (57) Weeks, J. D.; Chandler, D.; Andersen, H. C. Role of Repulsive Forces in Determining Equilibrium Structure of Simple Liquids. *J. Chem. Phys.* **1971**, *54*, 5237.
- (58) Caplan, D. A.; Subbotina, J. O.; Noskov, S. Y. Molecular Mechanism of Ion-Ion and Ion-Substrate Coupling in the Na<sup>+</sup>-Dependent Leucine Transporter LeuT. *Biophys. J.* **2008**, *95*, 4613.
- (59) Brooks, B. R.; Brucoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4*, 187.
- (60) 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. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998**, *102*, 3586.
- (61) Anisimov, V. M.; Lamoureux, G.; Vorobyov, I. V.; Huang, N.; Roux, B.; MacKerell, A. D. Determination of electrostatic parameters for a polarizable force field based on the classical Drude oscillator. *J. Chem. Theory Comput.* **2005**, *1*, 153.
- (62) Connolly, M. L. Solvent-Accessible Surfaces of Proteins and Nucleic-Acids. *Science* **1983**, *221*, 709.
- (63) Cantu, M. D.; Hillebrand, S.; Carrilho, E. Determination of the dissociation constants (pK(a)) of secondary and tertiary amines in organic media by capillary electrophoresis and their role in the electrophoretic mobility order inversion. *J. Chromatogr., A* **2005**, *1068*, 99.
- (64) Shalaeva, M.; Kenseth, J.; Lombardo, F.; Bastinz, A. Measurement of dissociation constants (pK(a) values) of organic compounds by multiplexed capillary electrophoresis using aqueous and cosolvent buffers. *J. Pharm. Sci.* **2008**, *97*, 2581.
- (65) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. The Weighted Histogram Analysis Method for Free-Energy Calculations on Biomolecules 0.1. The Method. *J. Comput. Chem.* **1992**, *13*, 1011.
- (66) Beglov, D.; Roux, B. Finite Representation of an Infinite Bulk System - Solvent Boundary Potential for Computer-Simulations. *J. Chem. Phys.* **1994**, *100*, 9050.
- (67) Mobley, D. L.; Dill, K. A. Binding of Small-Molecule Ligands to Proteins: "What You See" Is Not Always "What You Get. *Structure* **2009**, *17*, 489.

- (68) Swendsen, R. H.; Wang, J. S. Replica Monte-Carlo Simulation of Spin-Glasses. *Phys. Rev. Lett.* **1986**, *57*, 2607.
- (69) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141.
- (70) Li, J. B.; Zhu, T. H.; Cramer, C. J.; Truhlar, D. G. New class IV charge model for extracting accurate partial charges from wave functions. *J. Phys. Chem. A* **1998**, *102*, 1820.
- (71) Canutescu, A. A.; Shelenkov, A. A.; Dunbrack, R. L. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.* **2003**, *12*, 2001.
- (72) Cherubino, F.; Miszner, A.; Renna, M.; Sangaletti, R.; Giovannardi, S.; Bossi, E. GABA transporter lysine 448: a key residue for tricyclic antidepressants interaction. *Cell. Mol. Life Sci.* **2009**, *66*, 3797.

CT9006597

## Characterization of the BNNO Radical

Qianyi Cheng, Andrew C. Simmonett, Francesco A. Evangelista, Yukio Yamaguchi,  
and Henry F. Schaefer III\*

*Center for Computational Quantum Chemistry, 1004 Cedar Street, University of  
Georgia, Athens, Georgia 30602*

Received March 8, 2010

**Abstract:** The cyclic, trans, and cis BNNO molecules and the two isomerization reactions on their doublet electronic states potential energy surface (PES) are systematically investigated. Ab initio self-consistent field, complete active space self-consistent field, coupled cluster with single and double excitations (CCSD), and CCSD including perturbative triple excitations [CCSD(T)] quantum mechanical techniques are employed, in conjunction with Dunning's correlation consistent polarized valence basis sets (cc-pVXZ and aug-cc-pVXZ, where X = D, T, and Q). All stationary points located on the doublet PES lie within 19 kcal mol<sup>-1</sup> of the global minimum cyclic isomer at the aug-cc-pVQZ CCSD(T) level of theory. The cyclic and trans minima are separated by 2.4 kcal mol<sup>-1</sup> with an interconversion barrier (cyclic → TS2 → trans) of 18.3 kcal mol<sup>-1</sup>; the trans and cis isomers are separated by 10.4 kcal mol<sup>-1</sup> with a barrier (trans → TS1 → cis) of 10.4 kcal mol<sup>-1</sup>. The dissociation energies BNNO ( $\tilde{X}^2A'$ ) → B ( $^2P_u$ ) + NNO ( $\tilde{X}^1\Sigma^+$ ) for the cyclic, trans, and cis isomers are predicted to be 39.7, 37.3, and 27.0 kcal mol<sup>-1</sup>, respectively. The diatomic fragment dissociation energies BNNO ( $\tilde{X}^2A'$ ) → BN ( $X^3\Pi$ ) + NO ( $X^2\Sigma^+$ ) for the three isomers are determined to be 50.7, 48.4, and 38.0 kcal mol<sup>-1</sup>, respectively. Additionally, fundamental vibrational frequencies are computed for the cyclic and trans isomers through application of second-order vibrational perturbation theory (VPT2) at the cc-pCVTZ CCSD(T) level of theory. Comparison of the resulting vibrational frequencies and their isotopic shifts with those determined experimentally by Wang and Zhou yields the surprising result that the B ( $^2P_u$ ) + NNO ( $\tilde{X}^1\Sigma^+$ ) reaction leads to formation of the trans isomer. The latter structure is not the global minimum, rather the second lowest lying isomer. This apparent disparity is rationalized by detailed examination of the PES describing this reaction.

### Introduction

In the past few decades, boron nitrides have attracted much attention since they have various technical applications in nuclear technology and in the semiconductor and steel industries, taking advantage of their mechanical, thermal, and electrical properties as well as their chemical inertness.<sup>1,2</sup> For the amount of energy stored in a given system or region of space per unit volume, or per unit mass, boron is well-known for its high energetic density, among many kinds of propellant additives.<sup>3</sup> Therefore, boron has potential applications as an advanced fuel in propulsion systems.<sup>4</sup> During the burning of boron-containing propellants, some portions

of boron are oxidized to boron oxide releasing a large amount of energy, while some boron nitride (BN) is formed.<sup>5,6</sup>

As an important molecule in atmospheric chemistry, nitrous oxide (N<sub>2</sub>O) has also received considerable attention and interest. In the chemical industry, nitrous oxide is an effective oxidation agent.<sup>7–23</sup> Many experiments indicate that nitrous oxide is also important in the thermal decomposition of various propellants.<sup>3</sup> N<sub>2</sub>O is often used as a catalytic species for burn-rate modification of nitramine propellants as well.<sup>24,25</sup>

The reaction of boron and nitrous oxides and the resulting intermediate generation is an intriguing topic. In 2007, Wang and Zhou reported a combined matrix isolation infrared (IR) spectroscopic and theoretical study of the BNNO and AINNO

\* E-mail: sch@uga.edu.

molecules.<sup>26</sup> The BNNO and AlNNO molecules were prepared via the reactions of laser-evaporated boron and aluminum atoms with nitrous oxide (N<sub>2</sub>O) in solid argon and were identified on the basis of isotopically substituted IR absorptions as well as theoretical (density functional theory) calculations. From codeposition of laser-evaporated isotopic-enriched <sup>10</sup>B atoms with 0.5% N<sub>2</sub>O in argon matrix, a group of new IR absorptions at 1837.0, 1502.3, 838.2, and 633.8 cm<sup>-1</sup> were observed, along with strong N<sub>2</sub>O absorptions. These four absorptions were assigned to the B–N stretching (1837.0 cm<sup>-1</sup>), N–O stretching (1502.3 cm<sup>-1</sup>), N–N stretching (838.2 cm<sup>-1</sup>), and in-plane bending (633.8 cm<sup>-1</sup>) modes of the <sup>10</sup>BNNO molecule.<sup>26</sup> The experiment was repeated with naturally abundant boron atoms, yielding absorptions at 1795.7, 1500.3, 836.5, and 626.9 cm<sup>-1</sup> with IR intensities approximately four times stronger than the above-mentioned absorptions. The latter vibrational features were assigned to the corresponding modes of the <sup>11</sup>BNNO molecule. In order to confirm their findings, Wang and Zhou carried out B3LYP density functional theory (DFT) computations with the 6-311+G\* basis set; the BNNO molecule was predicted to have a <sup>2</sup>A' ground electronic state with a planar trans structure.

Wang, Li, Zhang, Sheng, and Yu reported a theoretical study of boron nitride (BN) generated from the boron atom and several nitrogen oxides.<sup>3</sup> BN is one of the products formed in the burning of a boron-containing propellant. Possible mechanisms for the reactions of boron and nitrogen oxides (NO, NO<sub>2</sub>, and N<sub>2</sub>O) were investigated using the G2-MP2 method. The reactions of the ground-state boron atom B (<sup>2</sup>P<sub>u</sub>) with nitrogen oxides were determined to be endothermic, while the reactions of an excited quartet state of the boron atom B (<sup>4</sup>P<sub>g</sub>) and nitrogen oxides are exothermic, and the BN product can be formed. For the BN formation reaction B (<sup>4</sup>P<sub>g</sub>) + N<sub>2</sub>O → BN + NO, two trans and two cis forms of the BNNO molecule were located on the quartet potential energy surface (PES). Among the four, one trans and one cis form were found as the reaction intermediates, and the other trans and cis BNNO structures were characterized as transition states from the intermediates to the final products (BN + NO).

In this study we make the first attempt to theoretically interpret the experimentally observed vibrational frequencies by explicitly considering the effects of anharmonicity. Furthermore, we extend the previous studies of the PES by employing significantly more reliable methodologies and reveal a previously neglected isomer which, surprisingly, is revealed to be the global minimum.

## Theoretical Methods

In this work, six correlation-consistent basis sets cc-pVXZ and aug-cc-pVXZ, where X = D, T, and Q, developed by Dunning and co-workers<sup>27,28</sup> were employed. Ab initio theoretical techniques included restricted open-shell Hartree–Fock (ROHF), unrestricted Hartree–Fock (UHF), complete active space self-consistent field (CASSCF),<sup>29,30</sup> spin-unrestricted coupled cluster with single and double excitations (UCCSD),<sup>31,32</sup> and UCCSD with perturbative triple excitations [UCCSD(T)].<sup>33–35</sup> For the unrestricted

coupled cluster computations, an ROHF reference wave function was used to control spin contamination. Computations were performed with the Molpro program suite,<sup>36</sup> the Mainz–Austin–Budapest (MAB) version of the ACESII program suite,<sup>37,38</sup> and PSI3<sup>39</sup> quantum chemistry packages.

The four core orbitals (1s-like orbitals of B, N, and O) were frozen in all correlated calculations. The *T*<sub>1</sub> diagnostic values<sup>40</sup> of the five stationary points are 0.027 (cyclic isomer), 0.022 (trans isomer), 0.035 (cis isomer), 0.032 (TS1), and 0.034 (TS2) at the cc-pVQZ UCCSD(T) optimized geometries. Analytic and numerical gradient methods were used to optimize geometries and to determine the dipole moments, harmonic vibrational frequencies, and associated IR intensities. Vibrational anharmonicities were computed by application of second-order perturbation theory<sup>41–48</sup> (VPT2) to the quartic force field. The Grendel<sup>49</sup> program was used to compute the force constants in internal coordinates, while Intder2005<sup>50–54</sup> was used to perform the nonlinear transformation of the force constants from the internal to Cartesian coordinates. The Anharm<sup>53,55</sup> program was adopted for the VPT2 analysis.

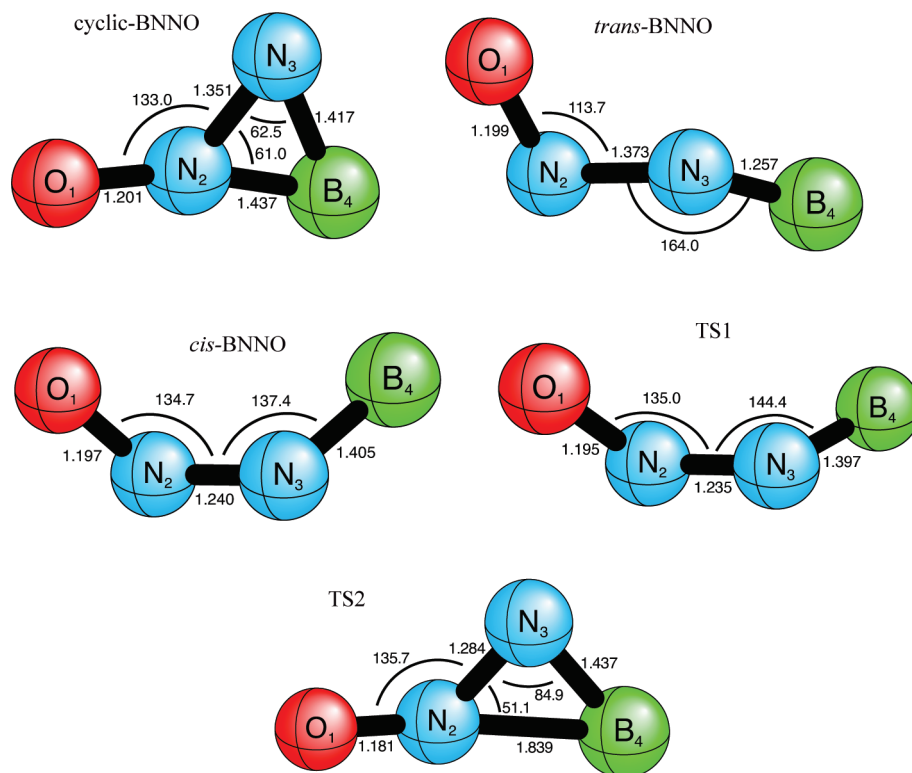
## Results and Discussion

**A. CASSCF Wave Functions.** The electron configurations of the linear NNO ( $\tilde{X}^1\Sigma^+$ ) and the cyclic, trans, and cis BNNO isomers as well as the two isomerization reaction transition states (TS1 and TS2) connecting them are shown in the Supporting Information, Table S1. The five highest-lying occupied molecular orbitals (MO) of the five stationary points of BNNO molecule and the linear NNO molecule are depicted in the Supporting Information, Figures S1–S6.<sup>57</sup>

In order to analyze correlation effects on the geometrical parameters and physical properties, full valence (19e/16MO) cc-pVQZ CASSCF wave functions were constructed for the five stationary points at the cc-pVQZ UCCSD(T) optimized geometries. There are 17 705 688 configuration state functions (CSFs) in *C*<sub>s</sub> point group symmetry. Furthermore, since multireference character might be present, given the moderately large *T*<sub>1</sub> diagnostics, we will examine the contributions of the reference wave function and the important excited configurations. Therefore, the CI coefficients based on natural orbitals (NOs), presented in the Supporting Information, Table S2, are employed in the following discussion.

For cyclic BNNO, the (2*a*'')<sup>2</sup> → (3*a*'')<sup>2</sup> double excitation provides the most significant correction to the reference configuration. For trans BNNO, the three most significant contributions to the CASSCF wave function come from the (2*a*'')<sup>2</sup> → (3*a*'')<sup>2</sup>, (11*a*'')<sup>2</sup> → (13*a*'')<sup>2</sup>, and (1*a*'')<sup>2</sup> → (4*a*'')<sup>2</sup> double excitations relative to the reference configuration. A major contribution to the CASSCF wave function for the cis isomer comes from the (13*a*'')<sup>2</sup> → (15*a*'')<sup>2</sup> double excitation. Similar to the cis isomer, the primary contribution to the CASSCF wave function of TS1 comes from the (13*a*'')<sup>2</sup> → (15*a*'')<sup>2</sup> double excitation, and the two important additional contributions are the (12*a*'')<sup>2</sup> → (16*a*'')<sup>2</sup> and (11*a*'')(13*a*'') → (14*a*'')(15*a*'') double excitations. It should be noted that the CI coefficient of the reference configuration for the TS1 transition state (*C*<sub>1</sub> = 0.912) is the same as the cis isomer





**Figure 1.** The optimized geometries (Å and °) of the five stationary point structures of BNNO at the aug-cc-pVQZ CCSD(T) level of theory.  $\tau$  (BNNO) for the cis BNNO isomer is 38.8°, and  $\tau$  for TS1 is 80.4°. The TS1 geometrical parameters are at the cc-pVQZ CCSD(T) level of theory.

but smaller than those for the cyclic and trans isomers. For TS2, three important double excitations,  $(2a'')^2 \rightarrow (3a'')^2$ ,  $(11a')^2(12a') \rightarrow (11a')(12a')(13a')$ , and  $(11a')^2 \rightarrow (13a')^2$  contribute to the CASSCF wave function. The CI coefficient of the reference configuration for the TS2 transition state ( $C_1 = 0.914$ ) is smaller than those of the cyclic and trans isomers, as a result of the elongated bonds in the transition state. Despite the significant presence of some excited configurations, the reference CI coefficients should be large enough in all cases for the single reference coupled cluster theory to be reliable.

**B. Geometries.** The optimized geometries for the five stationary points of the BNNO molecule are presented in Figure 1 and the Supporting Information, Table S3. In the following discussion we used the most reliable aug-cc-pVQZ CCSD(T) geometries for all species, with the exception of TS1, for which we encountered difficulties. Notwithstanding the change (0.5°) in  $\theta_e(\text{BNN})$  for the cis isomer, which has a very flat potential surface, it is evident from the Supporting Information, Table S3 that the aug-cc-pVQZ and cc-pVQZ geometries are nearly identical. In part for this reason, it was decided not to further pursue the aug-cc-pVQZ CCSD(T) geometry for TS1; instead we performed a single-point energy computation at this level of theory using the cc-pVQZ CCSD(T) geometry, anticipating negligible error in the resulting barrier height.

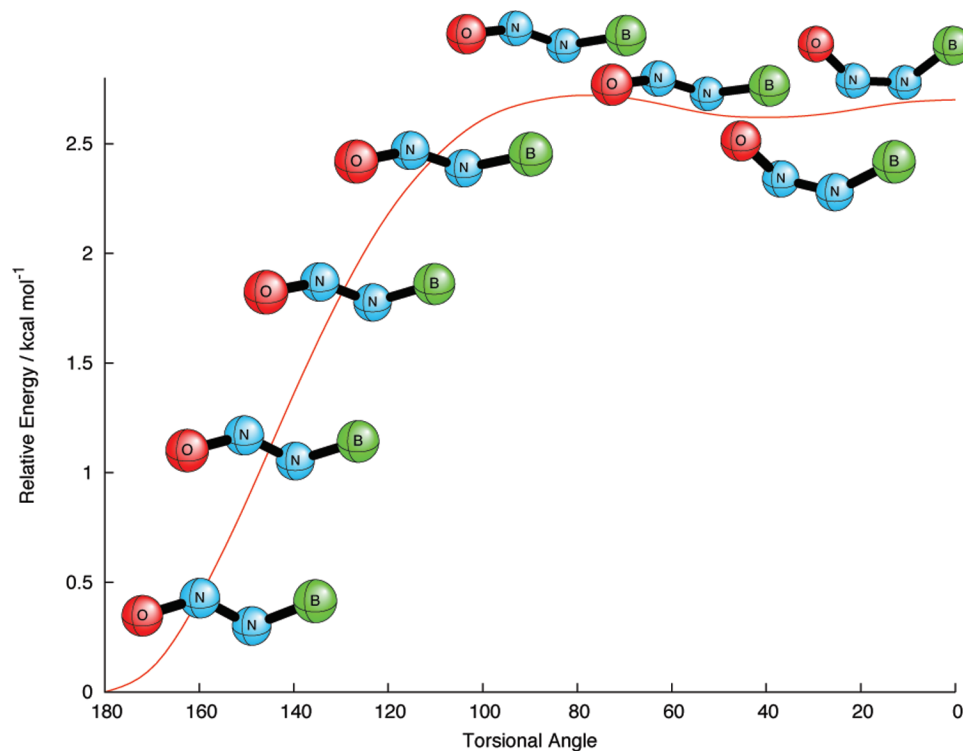
The  $r_e(\text{BN})$  bond distances in the five stationary points of the BNNO molecule are all predicted to be 1.417–1.437 Å, except for the trans minimum. These 1.4 Å values are very close to those of the boron- and nitrogen-containing three-membered rings from Richard and Ball's theoretical work<sup>58</sup>

[1.410 Å for both trans and cis diazaboridine and 1.420 Å for boradiazirine computed at the 6-31G(d,p) B3LYP level], in which B–N is a single bond. The trans BNNO isomer has the shortest BN-bond distance of 1.257 Å, which may be attributed to the two BN  $\pi$ -bonding (out-of-plane  $2a''$  and in-plane  $11a'$  MO) orbitals shown in the Supporting Information, Figure S2. This bond length is shorter than those for BN ( $X^3\Pi$ ) (1.330 Å), which is considered to have the character between a double and a triple bond, and BN ( $a^1\Sigma^+$ ) (1.277 Å),<sup>59</sup> which is considered as a triple bond.

The N–N bond distances of the cyclic and trans minima are 1.351 and 1.373 Å, respectively, much longer than those of isolated nitrous oxide (N–N triple bond<sup>56</sup> 1.129 Å) and diazene (N–N double bond 1.252 Å), much shorter than those of cis diazaboridine (1.585 Å)<sup>58</sup> and hydrazine (1.460 Å),<sup>60</sup> and close to that predicted for boradiazirine (1.300 Å)<sup>58</sup> at the B3LYP 6-31G(d,p) level. Therefore, the NN bonds in the cyclic and trans isomers fall between single and double bonds. For the cis minimum and the TS1 transition state, the N–N bond lengths are very close to diazene.

All of the O–N bond distances fall in the range 1.181–1.201 Å for the five stationary points. These values are much longer than that for ( $X^2\Pi$ ) diatomic nitric oxide (1.153 Å). Except for TS2, the O–N bond distances of the other four stationary points are also slightly longer than that for  $\text{N}_2\text{O}$  (1.188 Å).

The equilibrium bond angle  $\theta_e(\text{BNN})$  of the trans isomer is predicted to be the largest, around 164°, which suggests near  $sp$  hybridization for the B and N atoms. On the other hand, the bond angle  $\theta_e(\text{NNO})$  of the trans isomer is determined to be the smallest among the five stationary



**Figure 2.** Relaxed potential energy curve for BNNO, plotted as a function of the torsional angle  $\tau$  (BNNO), at the cc-pVTZ CCSD(T) level of theory.

points, about  $114^\circ$ , indicating something between  $sp^2$  and  $sp^3$  N and O hybridization. For the cis isomer, the bond angles  $\theta_e$  (BNN) and  $\theta_e$  (NNO) are predicted to be  $137.4^\circ$  and  $134.7^\circ$ , respectively. These geometrical features indicate a combination of  $sp$  and  $sp^2$  hybridization for the B and N atoms and the same type of hybridization for the O and N atoms in the cis isomer.

**C. Intrinsic Reaction Coordinate (IRC).** Intrinsic reaction coordinate (IRC) analyses<sup>61–64</sup> are commonly used to ascertain the nature of transition states; this requires locating a reaction coordinate, which is achieved by following appropriately mass-weighted energy gradients. The torsional motion that connects the cis and trans BNNO isomers has an extremely flat potential in the vicinity of the transition states, as exemplified by the small cis BNNO–TS1 separation of just  $0.06 \text{ kcal mol}^{-1}$ ; this makes gradient-following algorithms susceptible to numerical error. Instead, we manually varied the torsional angle, which is the primary contributor to the reaction coordinate, relaxing all other degrees of freedom to construct a potential energy curve at the cc-pVTZ CCSD(T) level of theory. This analysis, which is displayed in Figure 2, shows that the cis isomer does not reside in a deep enough potential well to be feasibly isolable and that its formation would immediately be followed by isomerization to trans BNNO. The region of Figure 2 around the  $C_1$  cis BNNO minimum reveals that equilibrium structure of this isomer is ill-defined, as isomerization between the two equivalent  $C_1$  minima occurs through a  $C_s$  symmetry transition state that is almost isoenergetic.

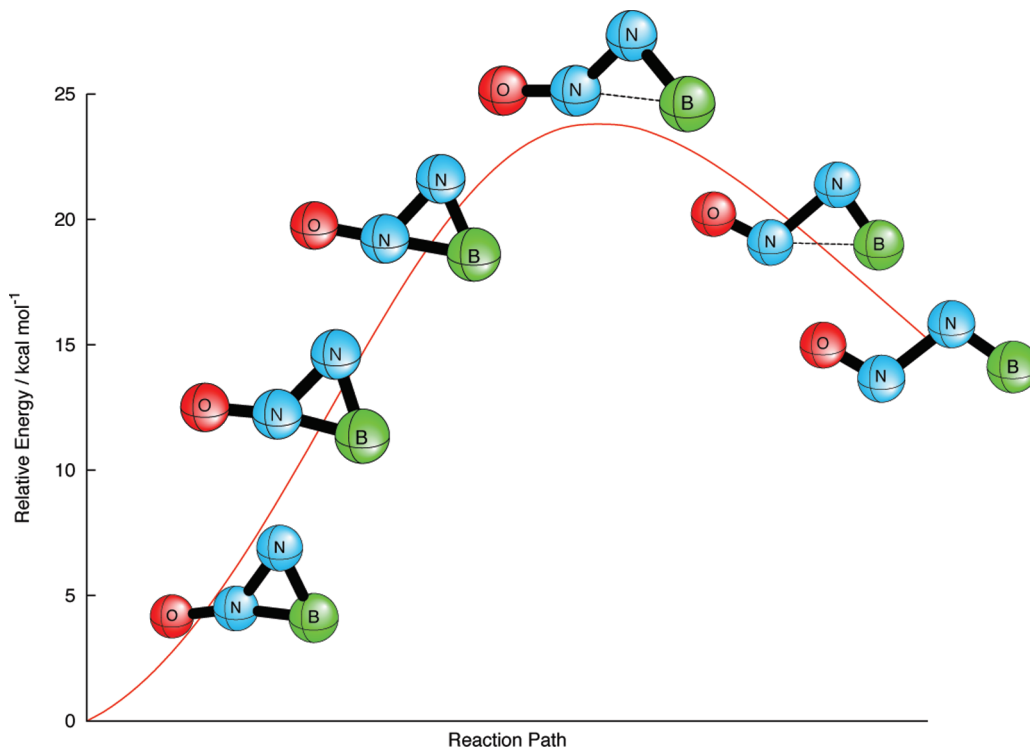
The transition state connecting the cyclic and trans BNNO minima is much more well-defined, and its cc-pVDZ MP2 IRC is plotted in Figure 3. For the forward reaction (cyclic  $\rightarrow$  TS2  $\rightarrow$  trans), the  $BN_3N_2$  bond angle (see Figure 1 for

atom numbering) of the cyclic isomer gradually opens up and the NN bond distance decreases toward TS2. At the transition state, the  $BN_2$  bond distance [ $1.839 \text{ \AA}$  with the aug-cc-pVQZ CCSD(T) method] is significantly elongated compared to that ( $1.437 \text{ \AA}$  with the same method) of the cyclic isomer. From the transition state (TS2) to the trans isomer, there is a cleavage of the  $BN_2$  bond, followed by shortening of the  $BN_3$  bond distance.

**D. Relative Energies.** The relative energies of the five stationary points are presented in Table 1. At the SCF level of theory, the trans isomer is predicted to be the energetically lowest-lying isomer. However, at the coupled cluster levels of theory, the cyclic isomer is found to be the global minimum on the ground-state surface. With the aug-cc-pVQZ CCSD(T) method, the trans BNNO structure is predicted to be higher in energy than the cyclic minimum, by  $3.5 \text{ kcal mol}^{-1}$  [ $2.4 \text{ kcal mol}^{-1}$  with zero-point vibrational energy (ZPVE) correction], but lower in energy than the cis isomer by  $10.6 \text{ kcal mol}^{-1}$  ( $10.4 \text{ kcal mol}^{-1}$  with ZPVE). The schematic PES at this level is shown in Figure 4.

The barrier height for the forward cyclic–trans isomerization reaction (cyclic  $\rightarrow$  TS2  $\rightarrow$  trans) is determined to be  $19.7 \text{ kcal mol}^{-1}$  ( $18.3 \text{ kcal mol}^{-1}$  with ZPVE), while the reaction barrier for the reverse reaction (trans  $\rightarrow$  TS2  $\rightarrow$  cyclic) is predicted to be  $16.2 \text{ kcal mol}^{-1}$  ( $15.9 \text{ kcal mol}^{-1}$  with ZPVE). Since the isomerization barrier heights are relatively high, the reaction may not happen at an appreciable rate in an argon matrix at 12K and would be highly dependent upon boron tunneling. This cyclic–trans isomerization reaction will be addressed again later in the manuscript.

The reaction barrier for the forward trans rotational (out-of-plane) isomerization reaction (trans  $\rightarrow$  TS1  $\rightarrow$  cis) was predicted to be  $10.8$  ( $10.4$ )  $\text{kcal mol}^{-1}$ . On the other hand,



**Figure 3.** Intrinsic reaction coordinate (IRC) for the cyclic–trans isomerization reaction of BNNO at the cc-pVDZ MP2 level of theory.

**Table 1.** Relative Energies of Five Stationary Points on the PES for the BNNO Molecule at SCF, CCSD, and CCSD(T) Levels of Theory<sup>a</sup>

level of theory	cyclic	trans	cis	TS1	TS2
cc-pVTZ SCF	0.00 (0.00)	−3.76 (−4.61)	17.94 (16.50)	19.50 (17.55)	18.43 (16.64)
aug-cc-pVTZ SCF	0.00 (0.00)	−3.35 (−4.19)	18.67 (17.22)	20.22 (18.25)	18.49 (16.70)
cc-pVQZ SCF	0.00 (0.00)	−3.59 (−4.44)	18.58 (17.13)	20.11 (18.14)	18.51 (16.72)
aug-cc-pVQZ SCF	0.00 (0.00)	−3.48 (−4.34)	18.76 (17.31)	20.28 (18.31)	18.54 (16.75)
cc-pVTZ CCSD	0.00 (0.00)	1.11 (0.15)	13.34 (11.50)	13.80 (12.31)	19.04 (17.14)
aug-cc-pVTZ CCSD	0.00 (0.00)	1.52 (0.58)	14.11 (12.24)	14.60 (13.10)	19.56 (17.66)
cc-pVQZ CCSD	0.00 (0.00)	1.48 (0.52)	14.64 (12.76)	15.09 (13.57)	20.21 (18.29)
aug-cc-pVQZ CCSD	0.00 (0.00)	1.52 (0.58)	14.84 (12.96)	15.30 (13.79)	20.38 (18.47)
cc-pVTZ CCSD(T)	0.00 (0.00)	3.22 (1.82)	12.81 (11.42)	12.91 (11.54)	18.49 (17.04)
aug-cc-pVTZ CCSD(T)	0.00 (0.00)	3.50 (2.33)	13.44 (12.00)	13.57 (12.16)	18.97 (17.53)
cc-pVQZ CCSD(T)	0.00 (0.00)	3.51 (2.33)	13.96 (12.61)	14.09 (12.65)	19.57 (18.09)
aug-cc-pVQZ CCSD(T)	0.00 (0.00)	3.50 (2.35)	14.10 (12.76)	14.25 (12.82) <sup>b</sup>	19.74 (18.27)

<sup>a</sup> Relative energies are in kcal mol<sup>−1</sup>. ZPVE corrected values are in parentheses. <sup>b</sup> ZPVE values computed at the cc-pVQZ CCSD(T) level of theory.

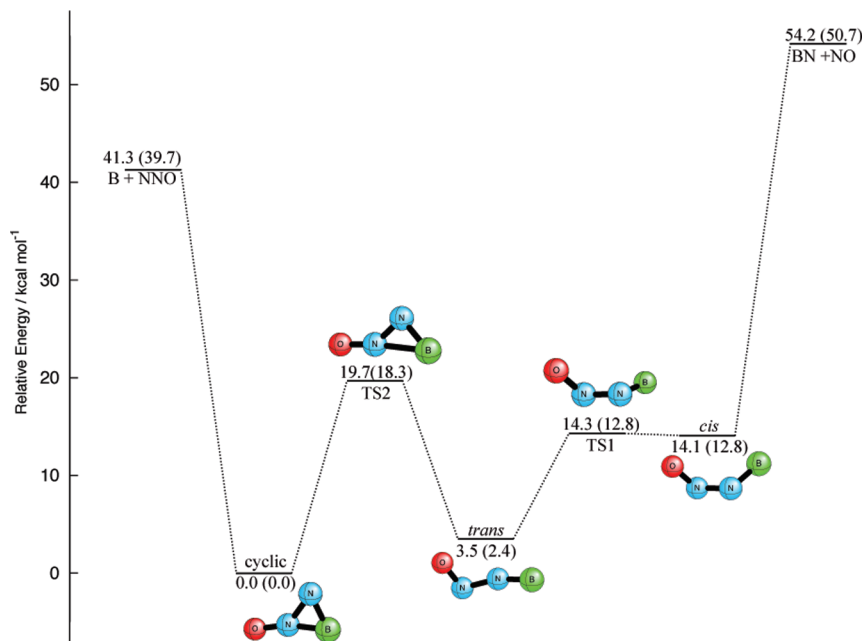
there is almost no barrier [0.2 kcal mol<sup>−1</sup> (0.0 kcal mol<sup>−1</sup> with ZPVE correction)] for the reverse isomerization reaction (cis → TS1 → trans). Consequently, the existence of the cis isomer in a solid argon matrix (at 12K) seems questionable.

**E. Dissociation Energies.** The two BNNO dissociation limits at the aug-cc-pVQZ CCSD(T) level of theory are shown schematically in Figure 4.

1.  $BNNO (\tilde{X}^2A') \rightarrow B (^2P_u) + NNO (\tilde{X}^1\Sigma^+)$ . The dissociation energies  $BNNO (\tilde{X}^2A') \rightarrow B (^2P_u) + NNO (\tilde{X}^1\Sigma^+)$  for the three BNNO isomers are presented in Table 2. With the aug-cc-pVQZ basis set the dissociation energy (ZPVE corrected values in parentheses) for the cyclic minimum is predicted to be 30.5 (28.5) (SCF), 38.3 (36.2) (CCSD), and 41.3 (39.7) kcal mol<sup>−1</sup> [CCSD(T)]. For the trans isomer, the three corresponding values are 34.0 (32.8) (SCF), 36.8 (35.6) (CCSD), and 37.8 (37.3) kcal mol<sup>−1</sup>

[CCSD(T)], while those for the cis isomer are 11.9 (11.3) (SCF), 23.6 (23.4) (CCSD), and 27.3 (27.0) kcal mol<sup>−1</sup> [CCSD(T)]. With inclusion of correlation effects, the dissociation energies increase relative to the SCF method by 11.2 (cyclic), 4.5 (trans), and 15.7 (cis) kcal mol<sup>−1</sup>, respectively. The cis isomer is more favored energetically by correlation effects compared those of the two dissociation products. It is seen that the  $B (^2P_u) + NNO (\tilde{X}^1\Sigma^+)$  dissociation pathways are endothermic for all three BNNO isomers (see Figure 4).

2.  $BNNO (\tilde{X}^2A') \rightarrow BN (X^3\Pi) + NO (X^2\Sigma^+)$ . The dissociation energies  $BNNO (\tilde{X}^2A') \rightarrow BN (X^3\Pi) + NO (X^2\Sigma^+)$  for the three BNNO isomers are reported in Table 2. The dissociation energy with the ZPVE correction for the cyclic minimum is 13.6 (SCF), 44.6 (CCSD), and 50.7 kcal mol<sup>−1</sup> [CCSD(T)]. For the trans isomer, the dissociation



**Figure 4.** Stationary points on the BNNO PES at the aug-cc-pVQZ CCSD(T) level of theory. Relative energies are in kcal mol<sup>-1</sup> (ZPVE corrected values in parentheses).

**Table 2.** Dissociation Energies of the BNNO ( $\tilde{X}^2A'$ )  $\rightarrow$  B ( $^2P_u$ ) + NNO ( $\tilde{X}^1\Sigma^+$ ) and BNNO ( $\tilde{X}^2A'$ )  $\rightarrow$  BN ( $X^3\Pi$ ) + NO ( $X^2\Sigma^+$ ) Channels at the SCF, CCSD, and CCSD(T) Levels of Theory with the aug-cc-pVQZ Basis Set<sup>a</sup>

level of theory	cyclic	trans	cis
BNNO ( $\tilde{X}^2A'$ ) $\rightarrow$ B ( $^2P_u$ ) + NNO ( $\tilde{X}^1\Sigma^+$ )			
aug-cc-pVQZ SCF	30.50 (28.45)	33.95 (32.75)	11.91 (11.32)
aug-cc-pVQZ CCSD	38.27 (36.17)	36.77 (35.60)	23.58 (23.35)
aug-cc-pVQZ CCSD(T)	41.27 (39.65)	37.81 (37.33)	27.30 (27.02)
BNNO ( $\tilde{X}^2A'$ ) $\rightarrow$ BN ( $X^3\Pi$ ) + NO ( $X^2\Sigma^+$ )			
aug-cc-pVQZ SCF	17.58 (13.55)	21.03 (17.86)	-1.00 (-3.57)
aug-cc-pVQZ CCSD	48.34 (44.61)	46.84 (44.04)	33.65 (31.79)
aug-cc-pVQZ CCSD(T)	54.22 (50.67)	50.75 (48.35)	40.25 (38.04)

<sup>a</sup> Dissociation energies in kcal mol<sup>-1</sup> and ZPVE corrected values are in parentheses.

energy is determined to be 17.9 (SCF), 44.0 (CCSD), and 48.4 kcal mol<sup>-1</sup> [CCSD(T)], whereas that for the cis isomer is -3.6 (SCF), 31.8 (CCSD), and 38.0 kcal mol<sup>-1</sup> [CCSD(T)] with the aug-cc-pVQZ basis set. For the three equilibrium structures, the increases of the dissociation energies with inclusion of correlation effects are 37.1 (cyclic), 30.5 (trans), and 41.6 (cis) kcal mol<sup>-1</sup>, respectively. These BN ( $X^3\Pi$ ) + NO ( $X^2\Sigma^+$ ) dissociation reactions are thermodynamically disfavored relative to the B ( $^2P_u$ ) + NNO ( $\tilde{X}^1\Sigma^+$ ) pathway discussed above (see Figure 4).

**F. Dipole Moments.** The dipole moments for the five stationary points are presented in the Supporting Information, Tables S6–S10. For the three equilibrium structures, the dipole moments are predicted to be 1.90 (cyclic), 2.27 (trans), and 1.16 (cis) debye at the aug-cc-pVTZ CCSD(T) (CCSD for cis) level of theory. The trans isomer has the largest dipole moment, with the expected sign <sup>+</sup>BNNO<sup>-</sup>.

**G. Harmonic Vibrational Frequencies.** The harmonic vibrational frequencies for the five stationary points of the BNNO molecule at the aug-cc-pVQZ CCSD(T) level of theory are reported in Table 3 and in the Supporting Information, Tables S6–S10. The four-highest frequencies

are predicted to be 1712, 1298, 1013, and 859 cm<sup>-1</sup> for the cyclic minimum and 1876, 1544, 874, and 646 cm<sup>-1</sup> for the trans isomer, while they are 1641, 1368, 1011, and 679 cm<sup>-1</sup> for the cis isomer.

The corresponding experimentally observed (fundamental) frequencies are 1795.7, 1500.3, 836.5, and 626.9 cm<sup>-1</sup> for the <sup>11</sup>B<sup>14</sup>N<sup>14</sup>O isotopologue.<sup>65</sup> Among three isomers, four vibrational frequencies of the <sup>11</sup>B<sup>14</sup>N<sup>14</sup>O isotopologue for the trans isomer are most consistent with Wang and Zhou's experimental values. A more detailed comparison of the theoretical fundamental frequencies with Wang and Zhou's experimental observations will be given in Section I.

**H. Infrared (IR) Intensities.** The IR intensities of the six vibrational modes for three equilibrium isomers are presented in the Supporting Information, Tables S6–S10. From the IR spectra of Wang and Zhou,<sup>65</sup> the IR intensities ( $I_s$ ) for the four observed modes for the trans isomer were concluded to be in the order  $I_2$  (NO stretching) >  $I_3$  (NN stretching) >  $I_1$  (BN stretching) >  $I_4$  (in-plane bending). This experimental ordering is well reproduced for the trans isomer using the CCSD(T) level of theory (see Table S7 in the Supporting Information), even within the double harmonic approximation.

**I. Anharmonic Vibrational Frequencies and Isotopic Shifts.** In Table 4, the fundamental vibrational frequencies for the <sup>10</sup>B<sup>14</sup>N<sup>14</sup>O isotopologue as well as the respective isotopic shifts of <sup>11</sup>B<sup>14</sup>N<sup>14</sup>O and <sup>10</sup>B<sup>15</sup>N<sup>15</sup>O are presented. The anharmonic vibrational frequencies are determined via VPT2 theory using our cc-pCVTZ CCSD(T) quartic force field. For the <sup>10</sup>B<sup>14</sup>N<sup>14</sup>O trans isotopologue, the deviations between theoretical harmonic and experimental fundamental frequencies of the four modes are +85, +50, +28, and +11 cm<sup>-1</sup>, respectively. On the other hand, the corresponding differences between theoretical anharmonic and experimental fundamental frequencies for the trans isomer are +36, +28, -6, and -15 cm<sup>-1</sup>. The improvement in the agreement with



**Table 3.** Theoretical Predictions of the Total Energy, Harmonic Vibrational Frequencies, And Zero-Point Vibrational Energy For the  ${}^2A'$  Cyclic, Trans, Cis, TS1, and TS2  ${}^{11}\text{B}{}^{14}\text{N}{}^{14}\text{NO}$  Molecule at the aug-cc-pVQZ CCSD(T) Level of Theory<sup>a</sup>

structure	total energy	$\omega_1(\text{a}')$	$\omega_2(\text{a}')$	$\omega_3(\text{a}')$	$\omega_4(\text{a}')$	$\omega_5(\text{a}')$	$\omega_6(\text{a}'')$	ZPVE
cyclic	-209.134170	1712	1298	1013	859	517	511	8.45
trans	-209.128591	1876	1544	874	646	149	137	7.30
cis	-209.111697	1641	1368	1011	679	189	85	7.11
TS1	-209.111462	(1675)	(1420)	(990)	(651)	(204)	(100 <i>i</i> )	(7.06)
TS2	-209.102719	1866	1234	1005	431	348	691 <i>i</i>	6.98

<sup>a</sup> Total energy is in hartree, harmonic vibrational frequency ( $\omega$ ) is in  $\text{cm}^{-1}$ , and ZPVE is in  $\text{kcal mol}^{-1}$ . For TS1, the total energy is the single-point energy with cc-pVQZ CCSD(T) optimized geometry, and harmonic vibrational frequencies and ZPVE are computed at the cc-pVQZ CCSD(T) level.

**Table 4.** Fundamental Vibrational Frequencies for the BNNO Molecule and the Corresponding Shifts upon Isotopic Substitution at the cc-pCVTZ CCSD(T) Level of Theory<sup>a</sup>

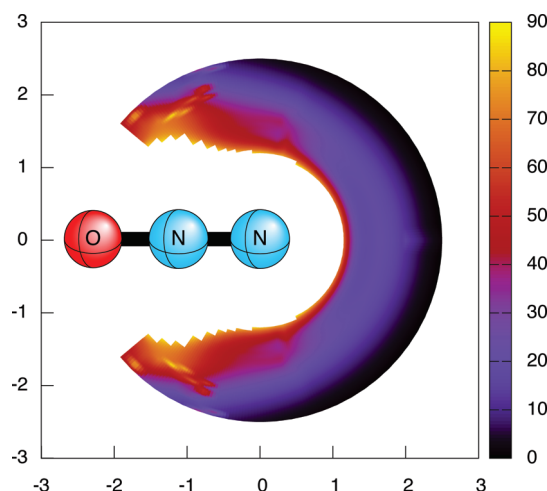
mode (sym.)	$\nu_{10}\text{B}{}^{14}\text{N}{}^{14}\text{NO}$	$\Delta(\nu_{11}\text{B}{}^{14}\text{N}{}^{14}\text{NO})$	$\Delta(\nu_{10}\text{B}{}^{15}\text{N}{}^{15}\text{NO})$
Experiment <sup>b</sup>			
B–N stretch	1837	41	28
N–O stretch	1502	2	26
N–N stretch	838	1	23
bending	634	7	9
Trans BNNO			
B–N stretch	1873	41	30
N–O stretch	1530	2	27
N–N stretch	832	1	22
bending	619	8	10
Cyclic BNNO			
B–N stretch	1305	28	19
N–O stretch	1684	15	34
N–N stretch	982	18	16
bending	830	6	16

<sup>a</sup> Fundamental vibrational frequencies are in  $\text{cm}^{-1}$ , and the corresponding shifts are denoted as  $\Delta$ . The experimental results are shown for comparison purposes. <sup>b</sup> Ref 26.

inclusion of theoretical anharmonic effects is evident. However, the disagreement between experiment is unusually large for such a reliable level of theory. Of course, the theoretical results are directly comparable only to gas-phase experiments, not matrix isolation results.

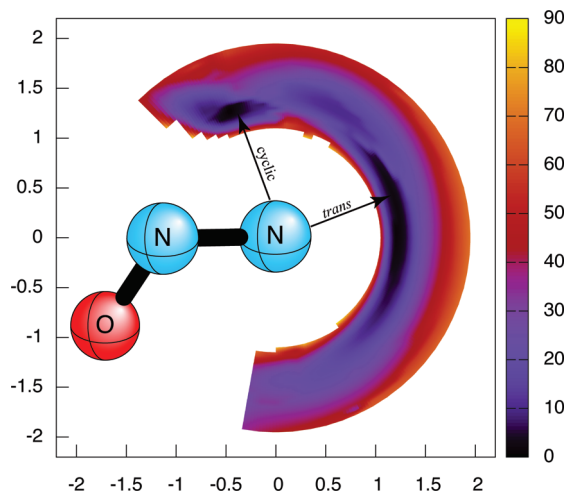
From the experimental observations, the  $1837\text{ cm}^{-1}$  transition exhibits the largest isotopic shift for  ${}^{11}\text{B}$  ( $41\text{ cm}^{-1}$ ) and  ${}^{15}\text{N}$  ( $28\text{ cm}^{-1}$ ), since it is primarily a B–N stretching mode. The  $1502\text{ cm}^{-1}$  frequency shows a very small  ${}^{11}\text{B}$  shift ( $2\text{ cm}^{-1}$ ) but quite a large nitrogen isotopic shift ( $26\text{ cm}^{-1}$ ), consistent with its assignment as the N–O stretching mode. The  $838\text{ cm}^{-1}$  absorption shows almost no change among boron isotopes but exhibits a large nitrogen isotopic shift ( $23\text{ cm}^{-1}$ ) and is, therefore, attributed to the N–N stretching mode.

For the global minimum cyclic isomer, the B–N stretching mode exhibits a large B isotopic shift ( $28\text{ cm}^{-1}$ ) and a N isotopic shift ( $19\text{ cm}^{-1}$ ). In the case of the trans isomer, the N–O and N–N stretching modes have almost no shift within B isotopes but quite large N isotopic shifts ( $27$  and  $22\text{ cm}^{-1}$  for the N–O and N–N stretches, respectively). Our frequency shifts upon isotopic substitution for the trans isomer are in good agreement with the experimentally observed values, in contrast to those for the cyclic global minimum; this indicates that the trans isomer was observed in the experiment.

**Figure 5.** PES (in  $\text{kcal mol}^{-1}$  and  $\text{\AA}$  units) describing the B ( ${}^2P_u$ ) + NNO ( $\tilde{X}{}^1\Sigma^+$ ) (linear) reaction at the cc-pVDZ CASSCF (7, 7) level of theory. See text for details.

**J. Trans or Cyclic Structure?** Our theoretical investigation clearly shows that the cyclic isomer is the global minimum, but the vibrational frequencies and isotopic shifts thereof provide compelling evidence for the experimental observation of the trans isomer. To gain some insight into the conformational preferences of the reaction, we constructed a two-dimensional energetic contour plot with respect to the boron-atom position, constraining the NNO moiety to its isolated ( $\tilde{X}{}^1\Sigma^+$ ) geometry and enforcing planarity. The cc-pVDZ CASSCF method with a (7, 7) active space was used in order to describe the various bonding schemes encountered on the resulting PES, which is shown in Figure 5. The area in the immediate vicinity of the molecule is repulsive within the constraints imposed, but crucially the region around the N terminus is less repulsive than that around the central nitrogen atom, favoring the formation of trans over cyclic BNNO. This feature may be explained from the NNO molecular orbitals shown in the Supporting Information, Figure S6. The  $7\sigma$  MO of NNO mainly consists of the lone-pair orbital of the terminal N atom. The electropositive B atom, therefore, may be prone to approach the electron-rich terminal N atom along the NNO molecular axis.

Clearly, as the reaction proceeds, the NNO angle must decrease, so an analogous plot was generated (Figure 6) with the NNO geometry chosen as [ $r_c(\text{NO}) = 1.30\text{ \AA}$ ,  $r_c(\text{NN}) = 1.20\text{ \AA}$ ,  $\theta_c = 123.0^\circ$ ] to represent a compromise between the NNO geometries adopted in the cyclic and trans isomers. The relaxation of the NNO unit changes the qualitative nature



**Figure 6.** PES (in kcal mol<sup>-1</sup> and Å units) describing the B (<sup>2</sup>P<sub>u</sub>) + NNO ( $\tilde{X}^1A'$ ) (bent) reaction at the cc-pVDZ CASSCF (7, 7) level of theory. Wells corresponding to cyclic and trans BNNO minima are labeled. See text for details.

of the potential, introducing two bound minima. The minimum corresponding to the cyclic structure, although deeper than the trans minimum, has a relatively small area, which translates into a relatively low capture cross section for boron atoms leading to cyclic BNNO formation. Although this analysis is quite crude, it offers insight into the basins of attraction on the B + NNO PES.

## Concluding Remarks

Ab initio molecular electronic structure theory has been employed in order to investigate the cyclic, trans, and cis isomers of BNNO and the two isomerization reactions connecting them. At our highest level of theory, aug-cc-pVQZ CCSD(T), the trans isomer was predicted to be 3.5 kcal mol<sup>-1</sup> (2.4 kcal mol<sup>-1</sup> with the ZPVE correction) higher than the cyclic minimum. The barrier height for the uphill isomerization reaction (cyclic → trans) is determined to be 19.7 (18.3) kcal mol<sup>-1</sup>. The trans and cis isomers are separated by 10.6 (10.4) kcal mol<sup>-1</sup> with a barrier (trans → TS1 → cis) of 10.8 (10.4) kcal mol<sup>-1</sup>, which indicates that the trans → cis isomerization reaction is unlikely to occur in an argon matrix. Theoretically computed harmonic and anharmonic vibrational frequencies and associated IR intensities are consistent with the experimental observation of trans BNNO in an argon matrix. The dissociation energies (with ZPVE corrections) associated with BNNO ( $\tilde{X}^2A'$ ) → B (<sup>2</sup>P<sub>u</sub>) + NNO ( $\tilde{X}^1\Sigma^+$ ) for the cyclic, trans, and cis isomers were predicted to be 39.7, 37.3, and 27.0 kcal mol<sup>-1</sup>, while the diatomic fragment dissociation energies BNNO ( $\tilde{X}^2A'$ ) → BN ( $X^3\Pi$ ) + NO ( $X^2\Sigma^+$ ) for the three isomers were determined to be 50.7, 48.4, and 38.0 kcal mol<sup>-1</sup>, respectively. Therefore, the three equilibrium structures are well below the dissociation limits to [B (<sup>2</sup>P<sub>u</sub>) + NNO ( $\tilde{X}^1\Sigma^+$ )] and [BN ( $X^3\Pi$ ) + NO ( $X^2\Sigma^+$ )]. There are no bonding regions with the NNO fragment constrained (linear) to its native geometry; its geometry must relax for the B (<sup>2</sup>P<sub>u</sub>) + NNO ( $\tilde{X}^1\Sigma^+$ ) association to proceed. Two distinct bonding regions are found on the relaxed (bent) surface. The first

leads to formation of the trans isomer, this region being much broader but less deep than that leading to formation of cyclic BNNO. The larger capture cross section due to the broader trans well appears to explain the experimental observation of only the higher energy trans isomer.

**Acknowledgment.** The authors would like to thank Dr. Partha P. Bera, Dr. Justin M. Turney, and Dr. Steven E. Wheeler for insightful discussions and technical expertise. This research was supported by the Department of Energy, Basic Energy Sciences, Division of Chemical Sciences, Fundamental Interactions Team, grant no. DEFG02-97-ER14748. This research used the resources of the National Energy Research Scientific Computing Center (NERSC), supported by the Office of Science of the U.S. Department of Energy under contract no. DE-AC02-05CH11231 and the VMD software developed by the Theoretical and Computational Biophysics Group in the Beckman Institute for Advanced Science and Technology at the University of Illinois at Urbana–Champaign.

**Supporting Information Available:** Electron configurations, CASSCF wave functions, molecular orbitals, dissociation energies, harmonic and fundamental vibrational frequencies, total energies, dipole moments, and IR intensities of the stationary points of BNNO at various levels of theory. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## References

- (1) Nohl, U.; Olbrich, G. *Gmelin Handbook of Inorganic Chemistry*; Springer: Berlin, Germany, 1988.
- (2) Paine, R. T.; Narula, C. K. *Chem. Rev.* **1990**, *90*, 73.
- (3) Wang, Z.; Li, S.; Zhang, L.; Sheng, Z.; Yu, S. *Propellants, Explos., Pyrotech.* **2004**, *29*, 160.
- (4) Kuo, K.; Pein, R. *Combustion of Boron-Based Solid Propellants and Solid Fuels*; CRC Press: Boca Raton, FL, 1993.
- (5) Chen, D. M.; Luh, S. P.; Liu, T. K.; Wu, G. K.; Perng, H. C. *Combustion of Boron-Based Solid Propellants and Solid Fuels* **1993**, 375.
- (6) Eckl, W.; Eisenreich, N.; Menke, K.; Rohe, T.; Weiser, V. Combustion phenomena of boron containing propellants. Proceedings of the 26th International Annual Conference of ICT, Karlsruhe, Federal Republic of Germany, July 4–7, 1995; p 70.
- (7) Ritter, D.; Weisshaar, J. C. *J. Phys. Chem.* **1990**, *94*, 4907.
- (8) Plane, J. M. C.; Rollason, R. J. *J. Chem. Soc., Faraday Trans.* **1996**, *92*, 4371.
- (9) Clemmer, D. E.; Honma, K.; Koyano, I. *J. Phys. Chem.* **1993**, *97*, 11480.
- (10) Matsui, R.; Senba, K.; Honma, K. *J. Phys. Chem.* **1997**, *101*, 179.
- (11) Campbell, M. L.; McClean, R. E. *J. Phys. Chem.* **1993**, *97*, 7942.
- (12) Campbell, M. L. *J. Chem. Phys.* **1996**, *104*, 7515.
- (13) Campbell, M. L.; Kölsch, E. J.; Hooper, K. L. *J. Phys. Chem.* **2000**, *104*, 11147.

- (14) Campbell, M. L. *J. Phys. Chem.* **2003**, *107*, 3048.
- (15) Armentrout, P. B.; Halle, L. F.; Beauchamp, J. L. *J. Chem. Phys.* **1982**, *76*, 2449.
- (16) Futерko, P. M.; Fontijn, A. *J. Chem. Phys.* **1991**, *95*, 8065.
- (17) Delabie, A.; Vinckier, C.; Flock, M.; Pierloot, K. *J. Phys. Chem.* **2001**, *105*, 5479.
- (18) Stirling, A. *J. Am. Chem. Soc.* **2002**, *124*, 4058.
- (19) Tishchenko, O.; Vinckier, C.; Nguyen, M. T. *J. Phys. Chem.* **2004**, *108*, 1268.
- (20) Tishchenko, O.; Vinckier, C.; Ceulemans, A.; Nguyen, M. T. *J. Phys. Chem.* **2005**, *109*, 6099.
- (21) Lavrov, V. V.; Blagojevic, V.; Koyanagi, G. K.; Orlova, G.; Bohme, D. K. *J. Phys. Chem.* **2004**, *108*, 5610.
- (22) Blagojevic, V.; Orlova, G.; Bohme, D. K. *J. Am. Chem. Soc.* **2005**, *127*, 3545.
- (23) Michelini, M. D. C.; Russo, N.; Alikhani, M. E.; Silvi, B. *J. Comput. Chem.* **2005**, *26*, 1284.
- (24) Palopoli, S. F.; Brill, T. B. *Combust. Flame* **1991**, *87*, 45.
- (25) Brill, T. B.; Brush, P. J.; Patil, D. G. *Combust. Flame* **1993**, *94*, 70.
- (26) Wang, G.; Zhou, M. *Chem. Phys. Lett.* **2007**, *342*, 90.
- (27) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007.
- (28) Woon, D. E.; Dunning, T. H. *J. Chem. Phys.* **1993**, *98*, 1358.
- (29) Knowles, P. J.; Werner, H.-J. *Chem. Phys. Lett.* **1985**, *115*, 259.
- (30) Werner, H.-J.; Knowles, P. J. *J. Chem. Phys.* **1985**, *82*, 5053.
- (31) Hampel, C.; Peterson, K. A.; Werner, H.-J. *Chem. Phys. Lett.* **1992**, *190*, 1.
- (32) Watts, J. D.; Gauss, J.; Bartlett, R. J. *Chem. Phys. Lett.* **1992**, *200*, 1.
- (33) Raghavachari, K.; Trucks, G. W.; Pople, J. A.; Head-Gordon, M. *Chem. Phys. Lett.* **1989**, *157*, 479.
- (34) Watts, J. D.; Gauss, J.; Bartlett, R. J. *J. Chem. Phys.* **1993**, *98*, 8718.
- (35) Stanton, J. F. *Chem. Phys. Lett.* **1997**, *281*, 130.
- (36) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M. *MOLPRO*, version 2006.1; University College Cardiff Consultants Limited: Wales, U.K., 2006.
- (37) Stanton, J. F.; Gauss, J.; Watts, J. D.; Lauderdale, W. J.; Bartlett, R. J. *Int. J. Quantum Chem.* **1992**, *44* (S26), 879.
- (38) Stanton, J. F.; Gauss, J.; Watts, J. D.; Szalay, P. G.; Bartlett, R. J.; with contributions from Auer, A. A.; Bernholdt, D. E.; Christiansen, O.; Harding, M. E.; Heckert, M.; Heun, O.; Huber, C.; Jonsson, D.; Jusélius, J.; Lauderdale, W. J.; Metzroth, T.; Michauk, C.; O'Neill, D. P.; Price, D. R.; Ruud, K.; Schiffmann, F.; Tajti, A.; Varner, M. E.; Vázquez, J. ACES II; Jürgen Gauss, John F. Stanton, and Peter G. Szalay: Mainz, Germany; Austin, TX; and Budapest 112, Hungary, 1992; <http://www.aces2.de>.
- (39) Crawford, T. D.; Sherrill, C. D.; Valeev, E. F.; Fermann, J. T.; King, R. A.; Leininger, M. L.; Brown, S. T.; Janssen, C. L.; Seidl, E. T.; Kenny, J. P.; Allen, W. D. *J. Comput. Chem.* **2007**, *28*, 1610.
- (40) Lee, T. J.; Taylor, P. R. *Int. J. Quantum Chem., Symp.* **1989**, *23*, 199.
- (41) East, A. L. L.; Johnson, C. S.; Allen, W. D. *J. Chem. Phys.* **1993**, *98*, 1299.
- (42) Nielsen, H. H. *Rev. Mod. Phys.* **1951**, *23*, 90.
- (43) Mills, I. M. In *Molecular Spectroscopy: Modern Research*; Rao, K. N., Mathews, C. W., Eds.; Academic Press: New York, 1972; p 115.
- (44) Watson, J. K. G. In *Vibrational Spectra and Structure*; Durig, J. R., Ed.; Elsevier: Amsterdam, The Netherlands, 1972; Vol. 6, p 1.
- (45) Papoušek, D.; Aliev, M. R. *Molecular Vibrational-Rotation Spectra*; Elsevier: Amsterdam, The Netherlands, 1982.
- (46) Clabo, D. A.; Allen, W. D.; Remington, R. B.; Yamaguchi, Y.; Schaefer, H. F. *Chem. Phys.* **1988**, *123*, 187.
- (47) Allen, W. D.; Yamaguchi, Y.; Császár, A. G.; Clabo, D. A.; Remington, R. B.; Schaefer, H. F. *Chem. Phys.* **1990**, *145*, 427.
- (48) Aarset, K.; Császár, A. G.; Sibert, E. L.; Allen, W. D.; Schaefer, H. F.; Klopper, W.; Noga, J. *J. Chem. Phys.* **2000**, *112*, 4053.
- (49) GRENDL is a program written by Jeremiah J. Wilke to perform general numerical differentiations to high orders of electronic structure data. Center for Computational Chemistry, University of Georgia: Athens, GA.
- (50) INTDER2005 is a general program developed by Wesley D. Allen and co-workers which performs various vibrational analyses and higher-order nonlinear transformations among force field representations. Center for Computational Chemistry, University of Georgia: Athens, GA.
- (51) Allen, W. D.; Császár, A. G. *J. Chem. Phys.* **1993**, *98*, 2983.
- (52) Allen, W. D.; Császár, A. G.; Szalay, V.; Mills, I. M. *Mol. Phys.* **1996**, *89*, 1213.
- (53) Sarka, K.; Demaison, J. In *Computational Molecular Spectroscopy*; Jensen, P., Bunker, P. R., Eds.; Wiley: Chichester, U.K., 2000; p 255.
- (54) Simmonett, A. C.; Evangelista, F. A.; Allen, W. D.; Schaefer, H. F. *J. Chem. Phys.* **2007**, *127*, 014306.
- (55) Yamaguchi, Y.; Schaefer, H. F. *ANHARM*, a FORTRAN program written for VPT2 analysis; Center for Computational Chemistry, University of Georgia: Athens, GA.
- (56) Wang, F.; Harcourt, R. D. *J. Phys. Chem. A* **2000**, *104*, 1304.
- (57) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33.
- (58) Richard, R. M.; Ball, D. W. *J. Mol. Struct.* **2007**, *806*, 113.
- (59) Karton, A.; Martin, J. M. L. *J. Chem. Phys.* **2006**, *125*, 144313.
- (60) MacKay, B. A.; Fryzuk, M. D. *Chem. Rev.* **2004**, *104*, 385.
- (61) Schaefer, H. F. *Chem. Britain* **1975**, *11*, 227.
- (62) Fukui, K. *Acc. Chem. Res.* **1981**, *14*, 363.
- (63) Schmidt, M. W.; Gordon, M. S.; Dupuis, M. *J. Am. Chem. Soc.* **1985**, *107*, 2585.
- (64) Garrett, B. C.; Redmon, M. J.; Steckler, R.; Truhlar, D. G.; Baldrige, K. K.; Bartol, D.; Schmidt, M. W.; Gordon, M. S. *J. Phys. Chem.* **1988**, *92*, 1476.
- (65) Wang, G.; Jin, X.; Chen, M.; Zhou, M. *Chem. Phys. Lett.* **2006**, *420*, 130.

## Erratum

**Implementation and Performance of DFT-D with Respect to Basis Set and Functional for Study of Dispersion Interactions in Nanoscale Aromatic Hydrocarbons.** [*J. Chem. Theory Comput.* 4, 2030–2048 (2008)]. By Roberto Peverati and Kim K. Baldridge\*.

Page 2046. A typographical error occurred in Table 9 of this manuscript with the  $s_6$  values of the B2PLYP DFT functional. Values reported as 1.55 should be 0.55, as also correctly reported in the original Figure 4, and the associated analysis in the main text.

**Table 9.** Summary of Density Functional Plus  $s_R/s_6$  Coefficient Combinations Proposed for a Variety of Basis Sets, As Determined from Predictions of S22 Complexes

DFT functional	basis set	$s_R$ value	optimized $s_6$ value	MAD (kcal/mol)
B97D	cc-pVDZ	1.1	1.00	1.075
	cc-pVDZ+CP	1.1	1.39	0.518
	cc-pVTZ	1.1	1.18	0.337
	cc-pVTZ+CP	1.1	1.41	0.454
	cc-pVQZ	1.1	1.26	0.330
	cc-pVQZ+CP	1.1	1.39	0.441
	TZV(2d,2p)	1.1	1.25	0.375
	TZV(2d,2p)+CP	1.1	1.38	0.425
	B3LYP	cc-pVDZ	1.1	0.73
cc-pVTZ		1.1	0.88	0.853
cc-pVQZ		1.1	0.96	0.612
PBE	cc-pVDZ	1.1	0.50	2.579
	cc-pVTZ	1.1	0.64	1.030
	cc-pVQZ	1.1	0.65	0.798
revPBE	cc-pVDZ	1.1	1.66	0.826
	cc-pVTZ	1.1	1.87	1.326
	cc-pVQZ	1.1	1.90	1.536
	cc-pVTZ (8–22)	1.1	1.87	0.393
	cc-pVQZ (8–22)	1.1	1.90	0.355
B2PLYP	cc-pVDZ	1.3	0.55	1.394
	cc-pVTZ	1.3	0.55	0.517

CT1002187

10.1021/ct1002187

Published on Web 05/13/2010