

JCTC

Journal of Chemical Theory and Computation

Theoretical Insights into Hydrogen Bonding and Its Influence on the Structural and Spectral Properties of Aquo Palladium(II) Complexes: cis -[(dppp)Pd(H₂O)₂]²⁺, cis -[(dppp)Pd(H₂O)(OSO₂CF₃)]⁺(OSO₂CF₃)⁻, and cis -[(dppp)Pd(H₂O)₂]²⁺(OSO₂CF₃)⁻₂

Guang-Jiu Zhao,[†] Ke-Li Han,^{*†} and Peter J. Stang^{*‡}

State Key Laboratory of Molecular Reaction Dynamics, Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian 116023, China, and Department of Chemistry, University of Utah, Salt Lake City, Utah 84112

Received May 1, 2009

Abstract: Density functional theory (DFT) and time-dependent density functional theory (TDDFT) methods have been performed to investigate the ground and excited states of aquo palladium(II) complexes cis -[(dppp)Pd(H₂O)₂]²⁺, cis -[(dppp)Pd(H₂O)(OSO₂CF₃)]⁺(OSO₂CF₃)⁻, and cis -[(dppp)Pd(H₂O)₂]²⁺(OSO₂CF₃)⁻₂. Insights into the influence of hydrogen bonding on the structural and spectral properties of these three aquo Pd(II) complexes are presented. The structures and the HOMO–LUMO energy gap of the three aquo Pd(II) complexes can be markedly influenced by hydrogen bonding interactions. Hydrogen bonds can also significantly influence their absorption spectra.

It is well-known that transition-metal solvent complexes, especially aquo transition-metal complexes, are of considerable experimental and theoretical interest because of their relevance in catalytic processes and as intermediates in reactions carried out in solution.¹ Recently, aquo palladium complexes have garnered increasing attention, and their synthesis, properties, and applications as catalysts and as synthetic intermediates have been widely studied.²

Stang et al. have reported that the reaction of cis -(1,3-bis(diphenylphosphino)propane)palladium(II) triflate with 1 and 2 equiv of water gives the corresponding monoquo and diaquo

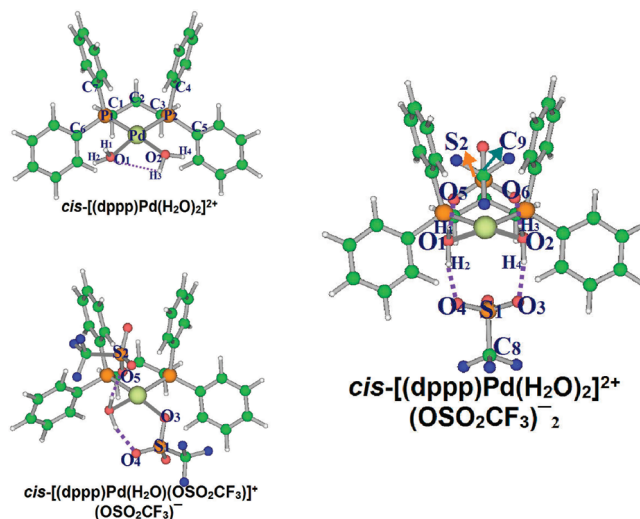


Figure 1. Conformational structures of the various aquo Pd(II) complexes.

cationic palladium(II) complexes cis -[(dppp)Pd(H₂O)(OSO₂CF₃)]⁺(OSO₂CF₃)⁻ and cis -[(dppp)Pd(H₂O)₂]²⁺(OSO₂CF₃)⁻₂, respectively, which are yellow microcrystalline solids.³ Moreover, the structures of monoquo and diaquo Pd(II) complexes have been determined by single-crystal X-ray crystallography.³

As is well-known, coordinated water forms hydrogen bonds with hydrogen bond acceptors in the presence of weakly coordinating counteranions (e.g., OTf⁻, ClO₄⁻, PF₆⁻, BF₄⁻, TsO⁻, NO₃⁻, etc.).² In the diaquo Pd(II) complex cis -[(dppp)Pd(H₂O)₂]²⁺(OSO₂CF₃)⁻₂ each coordinated water is doubly hydrogen-bonded to oxygen atoms of two different OSO₂CF₃⁻ anions (see Figure 1).³ In the monoquo Pd(II) complex cis -[(dppp)Pd(H₂O)(OSO₂CF₃)]⁺(OSO₂CF₃)⁻ the hydrogen atoms of coordinated water form two contacts, intramolecularly with an oxygen atom of the coordinated anion and intermolecularly with an oxygen atom of the OSO₂CF₃⁻ counterion (see Figure 1).³

Zhao et al. have extensively studied hydrogen bonding in both the ground and excited states of organic chromophores as well as its influence to their structures and dynamics in solution.⁴ However, hydrogen bonding in aquo transition-metal complexes, and its influence on the spectral properties of aquo transition-metal complexes, has been less investigated. Therefore, the density functional theory (DFT) and time-dependent density functional theory (TDDFT) methods have been applied to the study structural and spectral properties of these aquo Pd(II) complexes. All quantum chemical calculations have been performed using the Turbomole program suite along with the B3LYP hybrid functional due to its moderate efficiency and

* Corresponding author e-mail: klhan@dicp.ac.cn (K.-L.H.), stang@chem.utah.edu (P.J.S.).

[†] Chinese Academy of Sciences.

[‡] University of Utah.

Table 1. Calculated Important Bond Distances (Å) and Angles (deg) and the Corresponding X-ray Data³ (in Parentheses)

	[(dppp)Pd(H ₂ O) ₂] ²⁺	[(dppp)Pd(H ₂ O)(OSO ₂ CF ₃) ⁺ (OSO ₂ CF ₃) ⁻	[(dppp)Pd(H ₂ O) ₂] ²⁺ (OSO ₂ CF ₃) ⁻ ₂
Pd–O ₁	2.216	2.137 (2.106)	2.122 (2.135)
Pd–O ₂	2.189		2.122 (2.127)
Pd–O ₃		2.129 (2.159)	
Pd–P ₁	2.304	2.305 (2.237)	2.299 (2.231)
Pd–P ₂	2.304	2.305 (2.228)	2.299 (2.226)
P ₁ –C ₁	1.847	1.854 (1.812)	1.852 (1.804)
P ₁ –C ₆	1.831	1.832 (1.801)	1.831 (1.815)
P ₁ –C ₇	1.818	1.830 (1.805)	1.831 (1.799)
P ₂ –C ₃	1.846	1.854 (1.819)	1.852 (1.814)
P ₂ –C ₄	1.818	1.828 (1.815)	1.831 (1.803)
P ₂ –C ₅	1.831	1.832 (1.812)	1.831 (1.804)
O ₁ –H ₁	0.977	1.012 (0.71)	1.007 (0.83)
O ₁ –H ₂	0.981	0.997 (0.74)	1.001 (0.74)
O ₂ –H ₃	0.977		1.007 (0.76)
O ₂ –H ₄	0.979		1.001 (0.72)
S ₁ –O ₃		1.534 (1.469)	1.489 (1.431)
S ₁ –O ₄		1.480 (1.425)	1.489 (1.434)
S ₂ –O ₅		1.514 (1.426)	1.495 (1.448)
S ₂ –O ₆			1.495 (1.419)
O ₁ ···H ₃	2.679		
O ₃ ···H ₄			1.700 (1.93)
O ₄ ···H ₂		1.746 (2.04)	1.700 (2.00)
O ₅ ···H ₁			1.625 (1.83)
O ₆ ···H ₃			1.625 (2.03)
P ₁ –Pd–P ₂	91.21°	91.87° (90.98°)	93.91 (90.33)
O ₁ –Pd–O ₂	80.48°		93.07 (87.66)
O ₁ –Pd–O ₃		88.44° (86.1°)	
P ₁ –Pd–O ₁	95.51°	90.83° (93.2°)	86.38°(93.38°)
O ₂ –Pd–P ₂	92.80°		86.44°(89.42°)
O ₃ –Pd–P ₂		89.85° (89.5°)	
H ₁ –O ₁ –H ₂	107.4°	108.4° (107°)	113.2° (109°)
H ₃ –O ₂ –H ₄	108.7°		113.1° (114°)
O ₂ –H ₃ ···O ₁	89.62°		
O ₂ –H ₄ ···O ₃			161.2° (172°)
O ₁ –H ₂ ···O ₄		158.1° (157°)	161.2° (172°)
O ₁ –H ₁ ···O ₅		153.0° (153°)	164.6° (173°)
O ₂ –H ₃ ···O ₆			165.1° (163°)

accuracy and def-SV(P) basis set for nonmetal elements as well as def-TZVP and the pseudopotential def-ECP basis sets for Pd.⁵

Optimized conformational structures of the *cis*-[(dppp)-Pd(H₂O)₂]²⁺, *cis*-[(dppp)Pd(H₂O)(OSO₂CF₃)⁺(OSO₂CF₃)⁻], and *cis*-[(dppp)Pd(H₂O)₂]²⁺(OSO₂CF₃)⁻₂ complexes are presented in Figure 1, and some important computationally predicted bond distances and bond angles are listed in Table 1. All three aquo Pd(II) complexes have a square-planar coordination sphere. The calculated bond angles around the palladium atom of each complex range from 80.5°–95.5°, which are close to expected values of 90° and in accordance with X-ray data.³ Calculated palladium–phosphorus bond distances of 2.299–2.305 Å are in the normal range for *cis*-chelated phosphorus. Calculated Pd–OH₂ distances in *cis*-[(dppp)Pd(H₂O)₂]²⁺ are 2.216 and 2.189 Å, respectively, due to the different conformation of the two coordinated water molecules: one of which is perpendicular to the coordination plane while the other is roughly located within the plane. There is also a weak hydrogen bond formed between the two water molecules with a distance H₃···O₁ of 2.679 Å. From Table 1 it can be seen that all bond distances are in good agreement with the previously reported X-ray data³ with the exception of the bond distances to hydrogen atoms. The calculated bond distances of the O–H bonds are noticeably longer than those obtained from X-ray data. This is due to the well-known X-ray crystallographic underestimation of bond

distances to hydrogen atoms.³ As a corollary, the O···H hydrogen bond distances will be overestimated by X-ray crystallographic studies. Hence, hydrogen bond distances between coordinated water and OSO₂CF₃⁻ anions are calculated to be in the range of 1.625–1.746 Å, which are significantly shorter than the values from X-ray data.³ It is also important to note that Pd–OH₂ distances are shortened in *cis*-[(dppp)Pd-(H₂O)₂]²⁺(OSO₂CF₃)⁻₂ and *cis*-[(dppp)Pd(H₂O)(OSO₂CF₃)⁺-(OSO₂CF₃)⁻ due to the hydrogen bonding interactions between the coordinated water molecule(s) and OSO₂CF₃⁻ anions. At the same time, the bond angles of P–Pd–P and O–Pd–O are increased, and the bond angles of P–Pd–O are correspondingly decreased. In addition, the H–O–H angles of the coordinated water molecules are somewhat enlarged in the presence of the hydrogen bonding interactions.

Figure 2 shows the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) of the three aquo Pd(II) complexes studied computationally. It is noteworthy that the HOMO orbitals similarly have electron density localized on the *cis*-[(dppp)Pd(H₂O)₂]²⁺ moiety for all three aquo Pd(II) complexes. However, the electron density of the LUMO orbital has been delocalized from the *cis*-[(dppp)Pd(H₂O)₂]²⁺ moiety to the OSO₂CF₃⁻ anions for the hydrogen-bonded aquo Pd(II) complexes, which is believed to be induced by the hydrogen bonding interactions. Hence, LUMO orbitals of these aquo Pd(II) complexes can be influenced

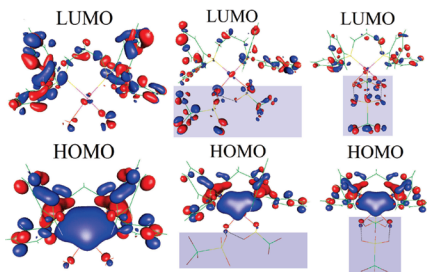


Figure 2. Frontier molecular orbitals (MOs) of the various aquo Pd(II) complexes.

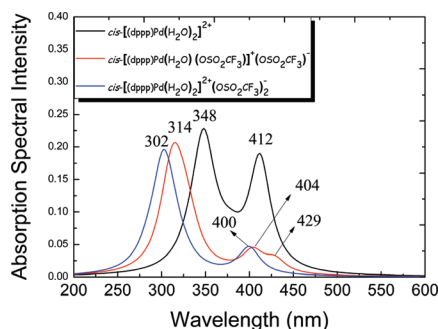


Figure 3. Calculated absorption spectra of the various aquo Pd(II) complexes. The absorption maxima values are labeled.

significantly by hydrogen bonding. Furthermore, the HOMO–LUMO energy gap of the *cis*-[(dppp)Pd(H₂O)₂]²⁺ complex increases from 3.72 to 4.40 eV by introducing two intermolecular hydrogen bonds in *cis*-[(dppp)Pd(H₂O)₂]²⁺(OSO₂CF₃)₂. The energy gap of *cis*-[(dppp)Pd(H₂O)(OSO₂CF₃)]⁺(OSO₂CF₃)⁻, on the other hand, decreases to 3.54 eV. The decrease of the HOMO–LUMO energy gap is ascribed to the substitution of one of the coordinated water molecules with an OSO₂CF₃⁻ anion.

The calculated absorption spectra of aquo Pd(II) complexes obtained using TDDFT methods are shown in Figure 3. The absorption spectrum of *cis*-[(dppp)Pd(H₂O)₂]²⁺ has a clear two-peak structure with maxima at 412 and 348 nm. The absorption maxima of *cis*-[(dppp)Pd(H₂O)₂]²⁺(OSO₂CF₃)₂ are blue-shifted to 400 and 302 nm, respectively. It is therefore observed that intermolecular hydrogen bonding in *cis*-[(dppp)Pd(H₂O)₂]²⁺(OSO₂CF₃)₂ can induce the absorption spectra shift to the blue, which is contrary to the spectral red-shift of most organic chromophores.⁴ This may be due to the electron delocalization of the LUMO orbital from the *cis*-[(dppp)Pd(H₂O)₂]²⁺ moiety to the two OSO₂CF₃⁻ anions induced by the intermolecular hydrogen bonding interactions. Consequently, the participation of the hydrogen-bonded OSO₂CF₃⁻ anion in LUMO orbital significantly increases the HOMO–LUMO energy gap of the *cis*-[(dppp)Pd(H₂O)₂]²⁺(OSO₂CF₃)₂ complex in comparison with that of *cis*-[(dppp)Pd(H₂O)₂]²⁺. Moreover, it is also demonstrated that the hydrogen-bonded OSO₂CF₃⁻ anion can shift the absorption spectra to the blue. For the case of *cis*-[(dppp)Pd(H₂O)(OSO₂CF₃)]⁺(OSO₂CF₃)⁻, it is interesting to note that a new red-shifted peak appears at 429 nm in addition to the blue-shifted absorption maxima at 404 and 314 nm. It is distinct that the hydrogen-bonded and coordinated OSO₂CF₃⁻ anions coexist in *cis*-[(dppp)Pd(H₂O)(OSO₂CF₃)]⁺(OSO₂CF₃)⁻ complex. Furthermore, both the hydrogen-bonded OSO₂CF₃⁻

anion and coordinated OSO₂CF₃⁻ anion participate in the electron delocalization of the LUMO orbital. Therefore, the blue-shifted absorption peaks may be dominantly due to the hydrogen-bonded OSO₂CF₃⁻ anion in *cis*-[(dppp)Pd(H₂O)(OSO₂CF₃)]⁺(OSO₂CF₃)⁻ complex. On the other hand, the new red-shifted peak at 429 nm is believed to have originated from the coordinated OSO₂CF₃⁻ anion. Moreover, the participation of the coordinated OSO₂CF₃⁻ anion in LUMO orbital can remarkably decrease the HOMO–LUMO energy gap of the *cis*-[(dppp)Pd(H₂O)(OSO₂CF₃)]⁺(OSO₂CF₃)⁻ complex. It is demonstrated that the coordinated OSO₂CF₃⁻ anion can shift the absorption spectra to the red, which is contrary to the hydrogen-bonded OSO₂CF₃⁻ anion. At the same time, the longest absorption wavelengths of these aquo Pd(II) complexes, which are 412, 400, and 429 nm for the *cis*-[(dppp)Pd(H₂O)₂]²⁺, *cis*-[(dppp)Pd(H₂O)(OSO₂CF₃)]⁺(OSO₂CF₃)⁻, and *cis*-[(dppp)Pd(H₂O)₂]²⁺(OSO₂CF₃)₂ complexes, are consistent with their corresponding HOMO–LUMO energy gaps of 3.72, 4.40, and 3.54 eV, respectively.

In summary, we have presented a theoretical investigation into the influence of hydrogen bonding on the structural and spectral properties of various aquo Pd(II) complexes. The structures and the HOMO–LUMO energy gaps of aquo Pd(II) complexes can be markedly changed by hydrogen bonding interactions. In addition, hydrogen bonds can also significantly influence absorption spectra. It has been demonstrated that the hydrogen-bonded OSO₂CF₃⁻ anion can shift the absorption spectra to the blue, while the coordinated OSO₂CF₃⁻ anion can shift the absorption spectra to the red. Further theoretical studies of the excited states of aquo transition-metal complexes will be performed in order to help understand the dynamic mechanism of hydrogen production by photocatalytic water splitting in many transition-metal complexes.⁶

Acknowledgment. K.L.H. thanks NKBRFSF (2007CB815202 and 2009CB220010) and NSFC (20833008) for financial support. P.J.S. thanks the NSF (CHE-0306720) for financial support.

Supporting Information Available: Cartesian coordinates and transition energies of optimized aquo Pd(II) complexes. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) (a) Beck, W.; Sünkel, K. *Chem. Rev.* **1988**, *88*, 1405. (b) Kubas, G. J.; Burns, C. J.; Khalsa, G. R. K.; Van Der Sluys, L. S.; Kiss, G.; Hoff, C. D. *Organometallics* **1992**, *11*, 3390. (c) Tamayo, A. B.; Alleyne, B. D.; Djurovich, P. I.; Lamansky, S.; Tsyba, I.; Ho, N. N.; Bau, R.; Thompson, M. E. *J. Am. Chem. Soc.* **2003**, *125*, 7377. (d) Alezra, V.; Bernardinelli, G.; Corminboeuf, C.; Frey, U.; Kundig, E. P.; Merbach, A. E.; Saudan, C. M.; Viton, F.; Weber, J. *J. Am. Chem. Soc.* **2004**, *126*, 4843. (e) Bunzli, J. C. G.; Piguet, C. *Chem. Soc. Rev.* **2005**, *34*, 1048. (f) Wik, B. J.; Lersch, M.; Krivokapic, A.; Tilset, M. *J. Am. Chem. Soc.* **2006**, *128*, 2682. (g) Beret, E. C.; Martinez, J. M.; Pappalardo, R. R.; Marcos, E. S.; Doltsinis, N. L.; Marx, D. *J. Chem. Theory Comput.* **2008**, *4*, 2108.
- (2) (a) Vicente, J.; Arcas, A. *Coord. Chem. Rev.* **2005**, *249*, 1135. (b) Jain, V. K.; Jain, L. *Coord. Chem. Rev.* **2005**, *249*, 3075. (c) Machotta, A. B.; Straub, B. F.; Oestreich, M. *J. Am. Chem. Soc.* **2007**, *129*, 13455. (d) Sakaguchi, S.; Yoo, K. S.; O'Neill, J.; Lee,

- J. H.; Stewart, T.; Jung, K. W. *Angew. Chem., Int. Ed.* **2008**, *47*, 9326. (e) Leininger, S.; Olenyuk, B.; Stang, P. J. *Chem. Rev.* **2000**, *100*, 853. (f) Lopez-Serrano, J.; Duckett, S. B.; Aiken, S.; Lenero, K. Q. A.; Drent, E.; Dunne, J. P.; Konya, D.; Whitwood, A. C. *J. Am. Chem. Soc.* **2007**, *129*, 6513.
- (3) Stang, P. J.; Cao, D. H.; Poulter, G. T.; Arif, A. M. *Organometallics* **1995**, *14*, 1110.
- (4) (a) Zhao, G.-J.; Han, K.-L. *J. Phys. Chem. A* **2007**, *111*, 2469. (b) Zhao, G.-J.; Liu, J.-Y.; Zhou, L.-C.; Han, K.-L. *J. Phys. Chem. B* **2007**, *111*, 8940. (c) Zhao, G.-J.; Han, K.-L. *ChemPhysChem* **2008**, *9*, 1842. (d) Zhao, G.-J.; Han, K.-L. *Biophys. J.* **2008**, *94*, 38.
- (5) (a) Furche, F.; Ahlrichs, R.; Wachsmann, C.; Weber, E.; Sobanski, A.; Vogtle, F.; Grimme, S. *J. Am. Chem. Soc.* **2000**, *122*, 1717. (b) Furche, F.; Ahlrichs, R. *J. Am. Chem. Soc.* **2002**, *124*, 3804. (c) Furche, F.; Ahlrichs, R. *J. Chem. Phys.* **2002**, *117*, 7433.
- (6) (a) Sakai, K.; Ozawa, H. *Coord. Chem. Rev.* **2007**, *251*, 2753. (b) Meyer, T. J. *Nature* **2008**, *451*, 778.

CT900216M

Effects of Long-Range Electrostatics on Time-Dependent Stokes Shift Calculations

Kristina E. Furse and Steven A. Corcelli*

Department of Chemistry and Biochemistry, 251 Nieuwland Science Hall, University of Notre Dame, Notre Dame, Indiana 46556

Received March 25, 2009

Abstract: Molecular dynamics simulations are essential to the correct interpretation of the response measured in time-dependent Stokes shift (TDSS) experiments of fluorescent probe molecules in biological environments. Within linear response theory, the TDSS response is the time correlation function of the fluctuations of $\Delta E(t)$, the difference between the solute environment interaction energy with the probe, modeled in both its electronically excited and ground states. $\Delta E(t)$ is dominated by electrostatic interactions between the environment and the ground- and excited-state charge distributions of the probe. The treatment of the long-ranged electrostatics in the calculation of the TDSS response in MD simulations is systematically investigated for three probes in aqueous solution: a model diatomic, coumarin 102, and Hoechst 33258. Nine different protocols for the treatment of the electrostatics were compared to particle mesh Ewald (PME), which was utilized as a reference standard. A computationally efficient pairwise alternative to PME, the damped shifted force method, was shown to reproduce the TDSS response calculated with PME for all three systems. In contrast, neglecting the role of the long-ranged electrostatics in the calculation of the TDSS response results in artifacts.

I. Introduction

There is currently great interest in using computational methods to interpret solvation dynamics experiments of fluorescent probe molecules in heterogeneous or confined environments.^{1–9} The typical practice is to collect an equilibrium MD simulation with the probe of interest modeled in its ground electronic state and then to postprocess the trajectory to calculate the solvation response, which describes the dynamics of the environment in the vicinity of the probe. Since the solvation response is generally short-ranged and is usually dominated by atoms within ~ 5 Å of the probe,^{4,10–12} it is not clear what role, if any, the handling of long-range electrostatics plays in computing the solvation response.

The complete electrostatic potential energy for a periodic system is the sum of the Coulomb interactions between all pairs of atomic partial charges, including their periodic replicas. Without further approximation, the direct calculation of this infinite sum is computationally intractable. An obvious

approach is the introduction of a spherical cutoff in which the Coulombic interactions are only calculated between pairs of charges whose spatial separation is less than a specified distance. However, this strategy fails because the electrostatic sum is conditionally convergent; the value of the sum as a function of the cutoff distance will not converge unless the portion of the system within the cutoff sphere is charge neutral, which is not generally the case. Moreover, cutoffs introduce a discontinuity into the energy and forces, which have some finite value within the cutoff radius that immediately drops to zero outside of the radius, leading to instabilities in simulations as atoms cross the boundary. Improper handling of interactions at a cutoff boundary can cause serious artifacts that dramatically affect the accuracy of calculations.

A number of different methods to handle long-range electrostatics have been developed that circumvent the difficulties posed by the conditional convergence of the electrostatic sum,^{13–15} including Ewald summation,^{16–19} implicit continuum dielectric,^{20,21} and fast multipole or reaction field approaches.^{14,19,22} In addition, for systems

* Corresponding author. E-mail: scorcell@nd.edu.

containing explicit dipoles, an explicit real-space lattice summation method has been developed by Ladd^{23,24} and applied to the calculation of static- and frequency- dependent dielectric constants^{25,26} as well as in the context of solvation dynamics.²⁷ Each of these methods has strengths and weaknesses that strike a different balance between physical accuracy and computational cost. Currently, the accepted methodology for the MD simulation of biomolecular systems is to treat the long-range electrostatics with an Ewald summation method, typically a computationally efficient variation called particle mesh Ewald (PME).^{16,17} The Ewald summation methods involve a division of the electrostatic sum into two absolutely convergent components: a real-space sum, representing the atoms within a specified cutoff radius, and a long-ranged reciprocal-space sum. While Ewald summation methods have proven to be extremely accurate, they are not ideal for postprocessing trajectories to calculate the solvation response for several reasons. First, they are nontrivial to implement when developing new methods, and they become computationally expensive for large systems. Most importantly, it is not clear how to decompose the collective solvation response to distinguish individual contributions in multicomponent systems, which represents a significant theoretical limitation.

Ideally, an electrostatics method for solvation dynamics calculations would involve a direct pairwise sum with minimal modifications to eliminate cutoff artifacts. One compelling option is based on two key observations described by Wolf et al.²⁸ First, electrostatic interactions are *effectively* short-ranged in condensed-phase systems. Therefore, for the larger boxes typical of modern biomolecular simulation with a proper treatment of the electrostatic sum, convergence can be achieved within the central unit cell and does not require inclusion of explicit periodic images. The second key observation is the critical importance of neutralization of charge within the cutoff radius, which renders the electrostatic sum absolutely convergent. This neutralization is achieved mathematically by shifting the Coulomb potential so that it effectively goes to zero at the cutoff radius. This operation also has the advantage of eliminating the discontinuities in the energies at the cutoff radius. The physical interpretation of this mathematical shift is a projection of image charges equal and opposite to the charges within the cutoff onto the cutoff sphere, achieving charge neutralization within the spherical cutoff surface.

Fennell and Gezelter²⁹ have extended and refined this pairwise alternative to the Ewald summation methods^{28–30} and extensively validated two variations of the shifted potentials against PME. MD simulations were conducted using seven different test systems, spanning a range from neutral liquids to ionic crystals. One of the new pairwise methods, damped shifted force (DSF), emerged as superior, accurately reproducing the energetic and dynamic results computed with PME. In addition to shifting the Coulomb potential to zero at the cutoff radius, DSF also shifts the Coulombic forces to zero at the cutoff, thereby incorporating the same image charges into the force calculation as well as eliminating discontinuities in the forces. The DSF method is also versatile; while it is fully appropriate for systems with

three-dimensional periodicity, it does not actually require periodic images so it can also be used for nonperiodic systems and for those with two-dimensional periodicity. The major advantage for postprocessing trajectories to calculate the solvation response is that DSF involves a simple pairwise sum that can be rapidly calculated and trivially decomposed to isolate component contributions (e.g., solvent, counterions, and biomolecules).

In order to investigate systematically the effect of different electrostatics protocols on solvation dynamics calculations, we have chosen three probe/solvent systems of increasing complexity: (1) a neutral, rigid model diatomic molecule in water, (2) a neutral, flexible probe, coumarin 102 (C102), in water, and (3) a positively charged, flexible probe, Hoechst 33258 (H33258), in water with a single Cl[−] counterion. The H33258^{3,31} and C102^{8,32–42} probes have been utilized for solvation dynamics studies of water at the DNA interface. After performing MD simulations for all three systems using the AMBER package with periodic boundary conditions and PME, we postprocessed each simulation using ten different solvation dynamics protocols that represent variations of three electrostatics methods: direct Coulombic sums, DSF, and PME as a reference standard. Because the strengths and weaknesses of the PME method have been extensively characterized in the literature,^{43–48} it provides an ideal benchmark to evaluate the accuracy of the other treatments of the long-ranged electrostatics in the context of computing time-dependent Stokes shift (TDSS) responses. We found that the handling of electrostatics is extremely important, even with a cutoff radius three to four times larger than the probe range. Not surprisingly, bare Coulombic cutoffs lead to artifacts in the solvation dynamics calculations for all three test systems. Incorporation of a standard switching function and group-based imaging was sufficient to resolve the problems in the two neutral systems. However, these simple strategies were not sufficient to resolve unexpected artifacts in the most complex test system, H33258. This is an important cautionary note since the H33258 system with its mobile counterion is more representative of the complex, multicomponent biological systems being studied today. All of the artifacts disappeared with the DSF methods, which appear to be a simple and efficient alternative to Ewald summation for solvation dynamics calculations.

II. Computational and Theoretical Methods

A. Solvation Dynamics Calculations. The central quantity in computational studies of solvation dynamics is the instantaneous solvation energy, $\Delta E(t)$, which is the difference in the solute–solvent interaction energy with the solute modeled in both its excited and ground electronic states. According to linear response theory, the solvation response or time-dependent Stokes shift (TDSS), $S(t)$, is equal to the equilibrium time correlation of fluctuations in ΔE :^{49,50}

$$S(t) \cong C_{0(1)}(t) = \frac{\langle \delta \Delta E(0) \delta \Delta E(t) \rangle_{0(1)}}{\langle |\delta \Delta E|^2 \rangle_{0(1)}} \quad (1)$$

Table 1. Abbreviations for Solvation Dynamics Methods

Ewald	particle mesh Ewald
DSF_00	damped shifted force, $\alpha = 0.0 \text{ \AA}^{-1}$
DSF_01	damped shifted force, $\alpha = 0.1 \text{ \AA}^{-1}$
DSF_02	damped shifted force, $\alpha = 0.2 \text{ \AA}^{-1}$
grpCut	pairwise Coulomb, group-based bare cutoff
grpCutSw	pairwise Coulomb, group-based cutoff with a switching function
grpNoCut	pairwise Coulomb, group-based imaging, no cutoff
atmCut	pairwise Coulomb, atom-based bare cutoff
atmCutSw	pairwise Coulomb, atom-based cutoff with a switching function
atmNoCut	pairwise Coulomb, atom-based imaging, no cutoff

where $\delta\Delta E(t) = \Delta E(t) - \langle \Delta E \rangle_{0(1)}$ and $\langle \dots \rangle_{0(1)}$ indicates an ensemble average in the either the ground or excited electronic states of the solute.

For the present calculations, which are based on methodology developed in numerous previous MD studies of solvation dynamics,^{4-7,27,49,51-68} it is assumed that the electronic excitation occurs so quickly that the geometry of the solute does not change (vertical transition). It is reasonable to model the transition classically as a redistribution of atomic partial charges. Within this description, $\Delta E(t)$ is easily computed for a given instantaneous configuration of the solvent as the difference in the solute-solvent electrostatic potential energy for the ground-state solute partial charges and for the excited-state partial charges, $\Delta E(t) = E_{\text{excited}}(t) - E_{\text{ground}}(t)$.⁴⁹

Direct Pairwise Coulomb Methods. For N_{solute} charged sites interacting with N_{solvent} charged solvent sites, the energy difference $\Delta E(t)$ is typically calculated as a pairwise Coulomb sum:

$$\Delta E(t) = \sum_{i=1}^{N_{\text{solute}}} \Delta q_i \sum_{j=1}^{N_{\text{solvent}}} \frac{q_j}{r} \quad (2)$$

where Δq_i is the difference in partial charge ($q_{\text{excited}} - q_{\text{ground}}$) on solute site i , q_j is the partial charge on solvent atom j , and r is the interatomic separation between solute atom i and solvent atom j at time t . We have considered six different variations on this basic approach (Table 1). Three involve calculating r as the minimum distance of individual solvent atoms to solute atoms (atom-based imaging and cutoffs), and three involve considering the solvent as intact molecules (group-based imaging and cutoffs). In the latter approach, an entire water molecule was considered to be within the cutoff if the oxygen atom fell within the specified radius. The cutoffs used in the present work were set equal to one-half the box length, approximately 14, 15, and 20 Å for the diatomic, C102 and H33258 systems, respectively.

For both the atom- and group-based methods, we performed three different solvation dynamics calculations: (1) using a bare cutoff, applying eq 2 only to atoms (or molecules) within a specified radius from solute atoms, (2) using a switching function, scaling the electrostatic interaction with solvent near the cutoff to eliminate the discontinuities in the electrostatic potential energy and forces, and (3) no cutoff, applying eq 2 to the entire system (not including periodic images). All are standard protocols that have been used to handle electrostatics, although bare atom-

based cutoffs have been largely phased out. We include them here for comparison since the DSF method utilizes atom-based imaging and cutoffs.

DSF. The cutoff neutralized damped shifted force (DSF) method is a variation on a simple pairwise Coulomb sum that eliminates artifacts that arise from truncating long-ranged electrostatic interactions with a cutoff, yielding results that are consistent with more complex Ewald summation methods.²⁸⁻³⁰ Using the DSF formalism, $\Delta E(t)$ is expressed as

$$\Delta E(t) = \sum_{i=1}^{N_{\text{solute}}} \Delta q_i \sum_{j=1}^{N_{\text{solvent}}} q_j \left[\frac{\text{erfc}(\alpha r)}{r} - \frac{\text{erfc}(\alpha R_c)}{R_c} + \left(\frac{\text{erfc}(\alpha R_c)}{R_c^2} + \frac{2\alpha \exp(-\alpha^2 R_c^2)}{\sqrt{\pi} R_c} \right) (r - R_c) \right] \quad (3)$$

where Δq_i is the difference in partial charge ($q_{\text{excited}} - q_{\text{ground}}$) on solute site i , q_j is the partial charge on solvent atom j , r is the interatomic separation between solute atom i and solvent atom j at time t , and R_c is the cutoff radius ($r \leq R_c$). The parameter α in the complementary error function controls the range of damping; its purpose is to accelerate convergence of the Coulombic sum. We evaluated three different values of α representing increasing degrees of damping: 0.0, 0.1, and 0.2 Å⁻¹. For all of the DSF calculations, simple atom-based imaging and cutoffs were used, analogous to atmCut. Preliminary tests confirmed that incorporating group-based imaging and/or switching functions had no significant effect on the DSF results.

Ewald Summation. For each trajectory frame, we calculated the energy of the entire system using the excited-state solute charges and the ground-state solute charges; their difference is $\Delta E(t)$. We used smaller cutoff radii for the Ewald summation solvation dynamics calculations since the cutoff simply represents the boundary between the real and reciprocal space terms: 10, 10, and 16 Å for the diatomic, C102 and H33258, respectively.

Component Contributions. For all three probe systems, the ΔE calculation methods sum the interaction energy of the solute with the water molecules within the cutoff, R_c (if present). For the simple diatomic probe, water is the only environmental component, $\Delta E = \Delta E_{\text{water}}$. For the two flexible fluorescent probes, C102 and H33258 (Figure 1), ΔE also includes the interaction of the solute with itself via the electrostatic terms in the solute intramolecular molecular mechanics potential energy. For the "self-solvation" component of ΔE , we adopted standard handling of the intramolecular electrostatic energy: nonbonded pair lists were generated for each solute atom so that 1-1, 1-2, and 1-3 interactions are ignored, and 1-4 interactions are scaled down by a factor of 1.2. In this notation, the numbers correspond to relative solute atom positions: 1-1 refers to interaction of an atom with itself, 1-2 refers to interaction of an atom with its bonded neighbors, 1-3 refers to interaction of an atom with atoms exactly 2 bonds away, and so on. These interactions are described by the bond, angle, and torsion terms in the force field, so the electrostatics term is eliminated or scaled.

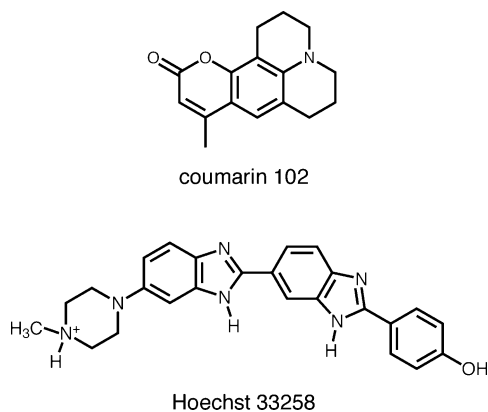


Figure 1. Structures of the fluorescent probes coumarin 102 and Hoechst 33258. The latter has a net charge of +1 due to the protonated piperazine group.

Since ΔE is calculated as a simple pairwise electrostatic sum in eqs 1 and 2, the results of the pairwise Coulombic and DSF solvation dynamics methods can be trivially decomposed into individual component contributions. For the C102 system, the instantaneous solvation energy has two components, water and solute conformation, $\Delta E = \Delta E_{\text{water}} + \Delta E_{\text{conf}}$. The positively charged H33258 system contains one additional component, the chloride counterion, $\Delta E = \Delta E_{\text{water}} + \Delta E_{\text{conf}} + \Delta E_{\text{ion}}$. Solvation response functions for each individual environmental component can be calculated by decomposing $C(t)$ into partial correlation functions:^{4,6}

$$C_{0(1)}^{\alpha}(t) = \frac{\langle \Delta E_{\alpha}(t) \Delta E(0) \rangle_{0(1)} - \langle \Delta E_{\alpha} \rangle_{0(1)} \langle \Delta E \rangle_{0(1)}}{\langle \Delta E(0) \Delta E(0) \rangle_{0(1)} - \langle \Delta E \rangle_{0(1)} \langle \Delta E \rangle_{0(1)}} \quad (4)$$

Decomposing the reciprocal space portion of the Ewald sum into component contributions is conceptually less straightforward and was not attempted here. In the absence of Ewald results, the full direct Coulomb sum with group-based imaging and no cutoff (grpNoCut) was used as a reference standard to evaluate the decompositions.

B. Parameter Development. The solutes studied here are not well represented by parameters available in the standard AMBER ff99 force field,⁶⁹ so new parameters were developed, including the creation of new atom types. As discussed in Section II-A, the atomic partial charges of the solute in its ground and excited electronic states are essential for computing solvation response functions. Force field parameters for the ground and excited states of all three probes were generated as follows.

Diatomic. The atoms of the diatomic molecule were given radii of 1.66 Å, a rigid bond length of 1.25 Å, and ground-state charges of 0 (similar to standard models of molecular oxygen). The hypothetical “excited state” was modeled with charges of +0.5 and −0.5 on the two atoms to create an instant dipole but maintain overall charge neutrality.

Coumarin 102. Coumarins are extremely well-studied fluorescent probes, so much of the structural and electronic information needed to calculate the missing force field parameters was available in the literature. The gas-phase minimum energy structure for C102, computed with density functional theory with the B3LYP^{70–72} functional and a 6-311G(d,p) basis set, was taken from Cave and Castner.⁷³

We took missing equilibrium bond lengths and angles from this minimum energy structure and interpolated the force constants using reference values in the ff99 force field.⁶⁹ The new equilibrium angles were adjusted to fit with existing ff99 parameters so that the interior angles of planar rings summed to the appropriate values (e.g., 720° for 6-membered rings). We set missing dihedral and improper torsion parameters equal to existing ff99 parameters that describe structurally similar rings found in amino and nucleic acids.

For the ground state, we calculated a molecular electrostatic potential over a grid of points using the optimized geometry and HF/6-311+G(d,p) basis set. We then fit the electrostatic potential to an atom-centered point charge model according to the RESP procedure.⁷⁴ Charges on topologically equivalent atoms were subsequently averaged (equivalenced) as in the standard AMBER charge fitting protocol. Basic CIS/6-31G(d) calculations proved insufficient to describe the excited state, giving no change in dipole moment. Ingrosso et al. calculated ground- and excited-state atomic charges for C153, which differs from C102 by a trifluoromethyl group and used them successfully to study solvation dynamics in acetonitrile.⁵⁵ C102 and C153 have similar spectral properties, including a similar change in dipole moment upon excitation.⁷³ We applied the atomic charge differences ($\Delta q = q_{\text{excited}} - q_{\text{ground}}$) from the C153 work to our ground-state charges to generate excited-state charges for C102. Classical dipole moment calculations on the resulting charge distributions using AMBER revealed that they were reasonable, giving an increase of 3.1 D upon excitation ($\mu_{\text{excited}} = 10.2$ D, $\mu_{\text{ground}} = 7.1$ D), which is consistent with the increase of 2.0 D ($\mu_{\text{excited}} = 9.2$ D, $\mu_{\text{ground}} = 7.2$ D) from other ab initio calculations on C102⁷³ as well as the range of values, 2.1–3.8 D, from microwave and solvatochromism experiments.^{75,76} All new C102 force field parameters, including charges, are included as Supporting Information.

Hoechst 33258. The fluorescent probe H33258, 2'-(4-hydroxyphenyl)-5-(4-methyl-1-piperazinyl)-2,5'-bibenzimidazole (Figure 1), is comprised of four structural units connected by three torsion angles (α , β , and γ), which are the key determinants of its conformation. It has greater inherent conformational flexibility than coumarin and required many ab initio calculations, including multiple torsion angle potential energy scans to describe its motion. It also exhibits more complex spectroscopy, including evidence of significant structural changes upon excitation, and proton transfer and rotation about the bisbenzimidazole axis, depending on pH and degree of confinement.^{31,77} The procedures we used to develop H33258 parameters are described in detail elsewhere, and the parameters are included as Supporting Information in that paper.¹⁰

C. Molecular Dynamics Simulations. MD simulations were performed using AMBER 9.0⁷⁸ with the AMBER ff99 force field⁶⁹ and the SPC/E water model.⁷⁹ The diatomic molecule and geometry-optimized conformations of C102 and H33258 were solvated in cubic periodic boxes with a minimum buffer of 10 Å from any solute atom to the closest box edge. A chloride counterion was added to the H33258 simulation to establish charge neutrality. The final unit cells were (1) diatomic with 673 water molecules: 2 021 total

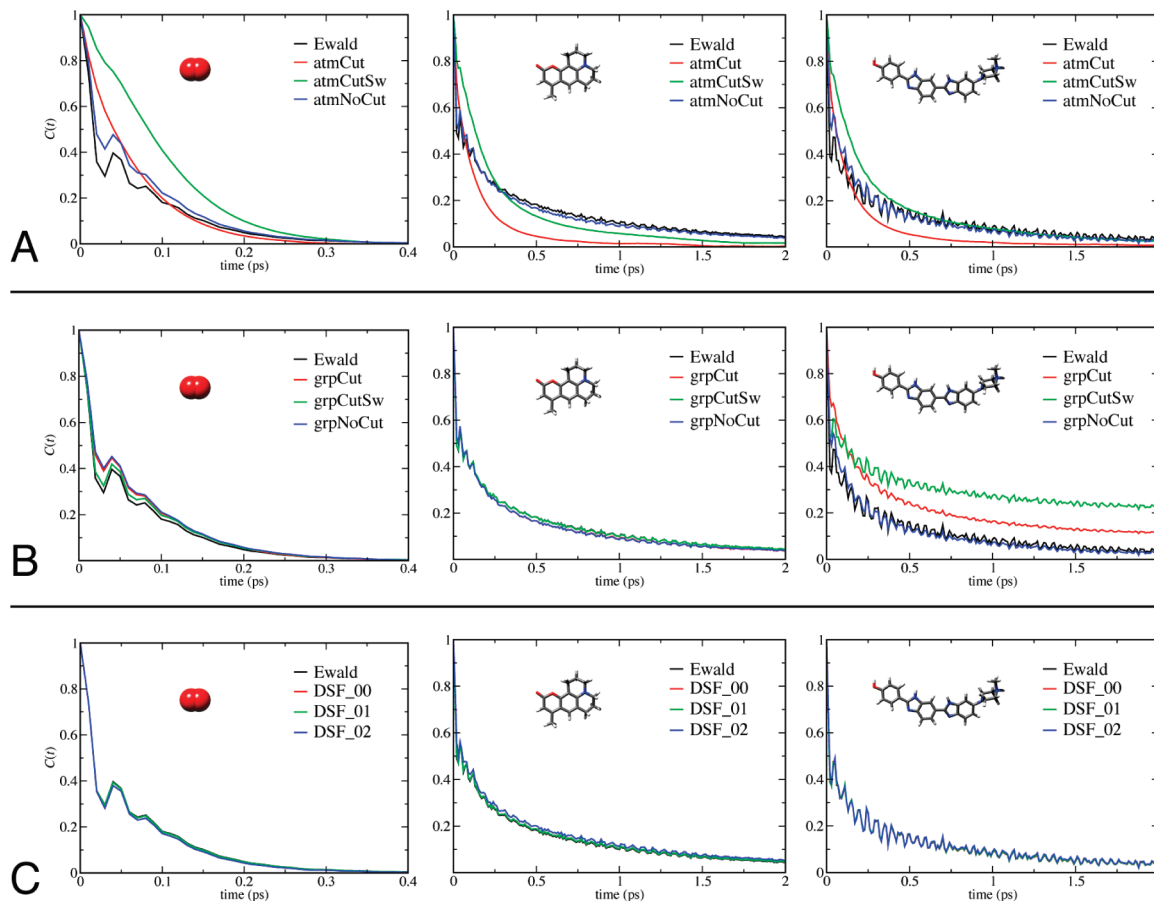


Figure 2. Comparison of time-dependent Stokes shift calculations using 10 different variations of direct Coulomb and DSF electrostatics methods. PME results (black curves) are included as a reference standard. Row A, direct Coulomb methods with atom-based imaging. Row B, direct Coulomb methods with group-based imaging. Row C, DSF methods. Each column represents a different solute in water: diatomic molecule, coumarin 102, and Hoechst 33258 with a Cl^- counterion.

atoms, (2) C102 with 973 water molecules: 2 955 total atoms, and (3) H33258 with Cl^- and 2 325 water molecules: 7 033 total atoms.

The same equilibration procedure was used for all three systems and is described in detail elsewhere.¹⁰ Production MD trajectories were performed in the NVE ensemble with the H33258 and C102 molecules fully flexible except for all covalent bonds containing hydrogen, which were fixed at equilibrium lengths using the SHAKE algorithm.⁸⁰ The diatomic was held rigid but could translate and rotate in the box, as could the fluorescent probes. A 2 fs integration time step was used, and configurations were collected every 10 fs for subsequent analysis. A PME summation method was used to compute long-range electrostatic energy and force corrections¹⁶ with a 9 Å real-space nonbonded cutoff. The data collection phase consisted of one 10 ns trajectory for the diatomic, four 6 ns trajectories (24 ns) for C102, and seven 1.5–6 ns trajectories (22.5 ns) for H33258.

III. Results and Discussion

A. Collective TDSS. Figure 2 shows the collective TDSS curves for all three test systems, using each of the nine pairwise electrostatics methods, compared to reference curves calculated with PME (black). Previous multiexponential fits to the H33258 DSF_02 data yielded time scales for the solvation response of H33258 free in aqueous solution (0.17

and 1.4 ps)¹⁰ and bound to DNA (1.5 and 20 ps)³ that were highly consistent with the results of ultrafast fluorescence experiments (0.2 and 1.2 ps, 1.4 and 19 ps, respectively),³¹ validating our simulations and general solvation dynamics methodology.

The TDSS curves generated with the DSF methods (Figure 2, row C) are nearly indistinguishable from the Ewald results for all three systems. Variation in the degree of dampening ($\alpha = 0.0, 0.1$ or 0.2) has little to no effect except for the C102 system, where the curve deviates slightly when $\alpha = 0.2$, suggesting minor overdamping. Motion of the solute itself (internal conformational changes, not overall translation and rotation) contributes oscillations that are proportional to the degree of solute flexibility.¹⁰ Internal motion of the C102 probe imparts minor oscillations and is responsible for a small portion ($\sim 3\%$) of the solvation response. Motion of the flexible H33258, which has four subunits that can rotate relative to one another, contributes more significant oscillations and is responsible for approximately one-third of the total solvation response. However, the short-lived oscillations in the computed rigid diatomic solvation response must have a different origin because there is no conformational freedom. Similar oscillations have been reported in numerous other studies of atomic and diatomic probes and have been attributed to coherent vibration of the solvation shell, an effect that is damped and far more subtle with polyatomic

probes.^{58,81} Interestingly, recent experimental studies indicate the presence of both types of oscillations for a quinolone probe: high-frequency oscillations corresponding to probe intramolecular motion in addition to a broader 170 fs oscillation feature attributed to coherent water motion.⁸² Related work in acetonitrile showed oscillations associated with solvent librational motion for certain nitrofluorenes but not for C153.⁸³ This is consistent with our simulations of the TDSS of C102 in water, which showed minor oscillations due to probe conformation but not coherent solvent motion. All of this suggests that the fast oscillatory components of the solvation response represent a complex interplay of probe and solvent structure and dynamics and warrant more study.

The results for the six direct Coulomb sum methods are much more variable than the DSF results. Atom-based imaging (with or without a switching function) shows the largest deviations from Ewald, giving featureless curves that exhibit the wrong time scales (Figure 2, row A). Reasonable accuracy is only recovered when the full interaction is calculated (atmNoCut), and even then the TDSS curves for the diatomic and H33258 systems deviate from the Ewald standard at early times. The DSF method also utilizes atom-based cutoffs, but the force shifting eliminates the artifacts seen with these direct Coulomb sums.

Group-based imaging is more robust, giving reasonable accuracy for the two neutral systems, the diatomic and C102 (Figure 2, row B). Interestingly, with group-based imaging, the TDSS curve calculated, using a cutoff with switching function (grpCutSw), is more consistent with Ewald than with that of the full interaction (grpNoCut). However, group-based imaging breaks down dramatically for the more complex H33258 system. The TDSS curves exhibit a long-time component and do not approach zero within 5 ps like the Ewald curve. This offset, which is even more pronounced when a switching function is added (grpCutSw), could be mistakenly interpreted as slow molecular motion in the system. Decomposition of the collective TDSS associated the long time component with movement of the ion across the cutoff boundary (see Supporting Information). The origin of this artifact in the H33258 system is explored further in the following section (Section III B).

B. H33258: Component Contributions. There are two primary factors that lead to the differences in the TDSS curves for the H33258 system (Figure 2, column 3): the fluctuation patterns of the solvent and ion components of the instantaneous solvation energy (ΔE_{water} and ΔE_{ion}) and their relative magnitudes. Since H33258 was always fully contained within the cutoff boundaries, the solute conformation (ΔE_{conf}) did not contribute to the observed differences and will not be considered here.

Atom-based cutoffs, which truncate solvent molecules at the cutoff radius, lead to dramatic overestimation of the solvent fluctuations (Figure 3, row A). Addition of a switching function reduces the fluctuations by an order of magnitude but does not completely resolve the problem. Calculating the full interaction (atmNoCut) restores the balance between the solvent and ion components, revealing important details like the inverse relationship between the two fluctuation patterns: as the ion contribution increases in

magnitude, the solvent contribution decreases. This relationship has been discussed in other computational studies.⁷

Keeping solvent molecules intact with group-based cutoffs eliminates the wild solvent fluctuations, but reveals a more subtle issue with the ion component (Figure 3, row B). As the ion moves away from the solute, its contribution to the instantaneous solvation energy should decrease in magnitude. When a bare cutoff is used, the interaction between the solute atoms and the ion instantly go to zero one by one: the ion passes out of range of the solute atoms that are farthest away, while it remains within the cutoff of nearer solute atoms. As a result, the ion sees a varying subset of the total change in solute charge distribution (which is no longer net neutral) as it moves across the cutoff, causing severe fluctuations in interaction energy where there should be a gradual decline (compare left and right panels of Figure 3, row B). Addition of a switching function reduces the magnitude of these spurious fluctuations but does not correct the fluctuation pattern, and therefore, still leads to an inaccurate TDSS. In fact, the false long-time component in the TDSS curve is even more pronounced (Figure 2, row B). Since the ion is monatomic, ΔE_{ion} is identical whether the cutoff is atom- or group-based, but with atom-based cutoffs, the anomalous behavior is masked by the solvent fluctuations and does not cause a long-time offset in the TDSS profile.

Although the DSF method utilizes an atom-based cutoff, it does not exhibit the same fluctuation artifacts seen with bare cutoffs and simple switching functions (Figure 3, row C). The solvent and ion fluctuation patterns for the DSF method are similar to the direct Coulomb sum with no cutoff, yielding similar TDSS profiles. The one caveat with the DSF methods is the slight reduction in the magnitude of the fluctuations compared to the results for the full interaction, proportional to the degree of damping. Since the reduction affects both solvent and ion components, the balance between components is maintained, but care should be taken with any calculation that involves the absolute magnitude of the fluctuations (Section III C).

C. Static Stokes Shift. In addition to the dynamic aspects of the solvation response, researchers are also interested in calculating the magnitude of the static Stokes shift. We calculated the static Stokes shift for each of the systems with all 10 long-range electrostatics methods (Table 2). The large variation in the magnitude of fluctuations in ΔE for the direct Coulomb methods (Figure 3) leads to a broad range of estimates of the static Stokes shift. Bare atom-based cutoffs are the worst, overestimating the static Stokes shift by 500- to 2 500-fold. Switching functions and group-based cutoffs reduce the estimates significantly but do not completely alleviate the problem. Even when no cutoff is used, the static Stokes shift is still overestimated compared to Ewald summation. In contrast, the DSF methods give results for the static Stokes shift that are most similar to the Ewald benchmarks. For example, with a damping parameter of 0.1 \AA^{-1} , DSF produced static Stokes shifts that were in error by just 3.3, 1.8, and 7.5% for the model diatomic, C102, and H33258 in aqueous solution, respectively.

To lend perspective to the comparisons in Table 2, we have included experimental estimates of the static Stokes

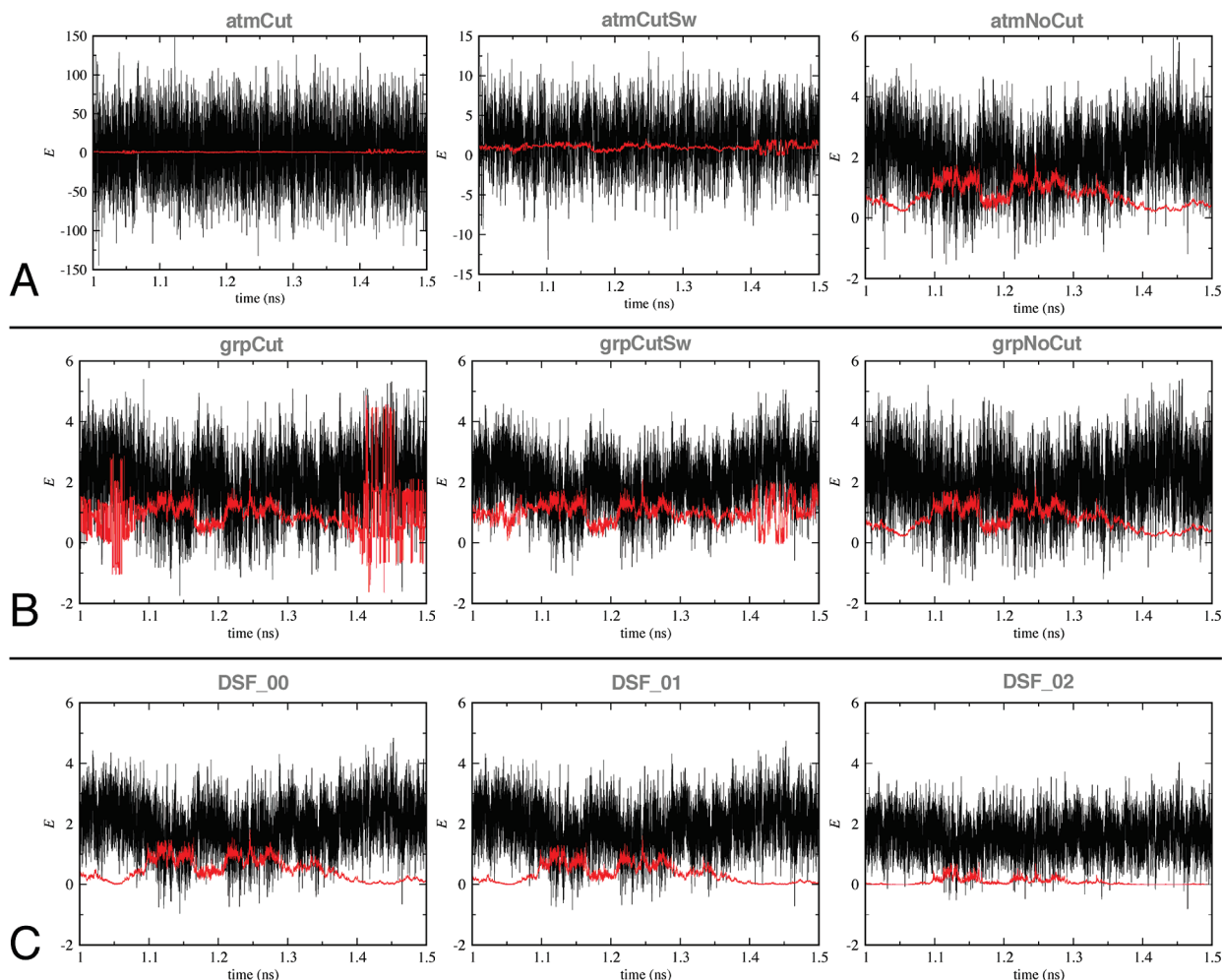


Figure 3. Individual solvent (black) and ion (red) contributions to the instantaneous solvation response, $\Delta E_{\text{water}}(t)$ and $\Delta E_{\text{ion}}(t)$ in kcal/mol, calculated for simulations of the fluorescent dye Hoechst 33258 in water with a Cl^- counterion. Each plot represents solvation dynamics results using a different electrostatics method: Row A, direct Coulomb with atom-based imaging: bare cutoff (left), switching function (center), and no cutoff (right). Row B, direct Coulomb with group-based imaging: bare cutoff (left), switching function (center), and no cutoff (right). Row C, DSF: $\alpha = 0.0$ (left), $\alpha = 0.1$ (center), and $\alpha = 0.2$ (right).

Table 2. Static Stokes Shift Comparison^a

method	diatomic	C102	H33258
Ewald	9.0	6.0	0.80
DSF_00	8.5	5.5	0.75
DSF_01	8.7	5.7	0.74
DSF_02	8.0	4.2	0.58
grpCut	12.5	7.2	2.43
grpCutSw	9.7	6.3	1.10
grpNoCut	13.1	7.4	1.28
atmCut	6 622	2 669	1 799
atmCutSw	425	109	13.99
atmNoCut	12.1	7.2	1.17
experiment	—	5.6 ⁸⁴	9.1 ³¹

^a $\Delta\Delta E = (\langle \delta \Delta E^2 \rangle_0) / k_{\text{B}} T$, results given in kcal/mol.

shift for C102⁸⁴ and H33258³¹ in water. Since C102 is only sparingly soluble in water, our result is a prediction, and the experimental result is shown for a related coumarin molecule, C343. Cave and Castner⁷³ have established that these molecules (C102 and C343) have almost identical spectral properties, including a similar change in dipole moment upon electronic excitation. Therefore, C102 would be expected to have a similar static Stokes shift as C343. Our Ewald, DSF

and grpCutSw calculations are all consistent with the experimental results for C102 but significantly underestimate the shift for H33258. To elucidate the origin of this discrepancy, it is important to emphasize the intrinsic difficulties associated with accurate computations of the static Stokes shift using the same classical models that are successful in computing the TDSS. Obtaining an accurate static Stokes shift requires knowledge of the average ground and excited electronic charge distribution of the solute molecules in their condensed-phase environment. However, in modeling these systems classically, the required charge distributions were calculated using quantum chemical calculations performed with the solute molecules in the gas phase. Owing to linear response theory, the TDSS, which is primarily dictated by solvent properties and the overall spatial variation in the electric field the solvent experiences upon charge rearrangement in the solute,^{58,81} is robust with respect to the overall accuracy of the ground- and excited-state charge distributions. For example, despite the inaccuracy of the static Stokes shift for H33258 in water, the computed time scales for the TDSS in aqueous solution (0.17 and 1.4 ps)¹⁰ and bound to DNA (1.5 and 20 ps)³ are in excellent

agreement with experiment (0.2 and 1.2 ps, 1.4 and 19 ps, respectively).³¹ It is a success that the DSF computed static Stokes shifts are in agreement with the Ewald benchmarks, even if the agreement with experiment is inadequate. MD calculations of more accurate static Stokes shifts will require continued advances in electronic structure that yield accurate excited-state charge distributions in the condensed phase.

IV. Summary and Conclusions

In this work, we have studied three systems of increasing complexity to determine the effect of different electrostatics methods on solvation dynamics calculations. While it is generally true that the solvation response is dominated by atoms within ~ 5 Å of the probe, we have found that improper handling of long-range electrostatics and cutoffs can introduce serious artifacts that could lead to misinterpretation of solvation dynamics calculations. For neutral systems with homogeneous solvent, a simple direct Coulomb sum with a group-based cutoff and switching function was sufficient to achieve reasonably accurate results for both the TDSS and the static Stokes shift. However, for the charged solute with a mobile counterion, this method broke down, giving an artificial long-time component. Our results suggest that for more complex, multicomponent systems with mobile charged groups, a method such as DSF or Ewald should be used to eliminate cutoff artifacts and address long-range electrostatics.

DSF has several important advantages over Ewald-based methods for these types of calculations, including ease of implementation, speed, and trivial decomposition into component contributions. If the research question relates to dynamic properties, such as the TDSS, then DSF methods rival Ewald in terms of accuracy for a fraction of the cost. The one potential disadvantage of DSF-based methods is that excessive damping reduces the magnitude of fluctuations in the instantaneous solvation energy, leading to an underestimation of the static Stokes shift. Careful choice of a damping parameter can minimize this reduction. DSF gave results for the static and dynamics Stokes shift that were the most consistent with Ewald of all of the systems studied here.

Acknowledgment. SAC gratefully acknowledges the support of the National Science Foundation (CHE-0845736), the Northwest Indiana Computational Grid, and the Notre Dame Center for Research Computing. The authors also thank Professor J. Daniel Gezelter and Dr. Charles F. Vardeman II for helpful discussions.

Supporting Information Available: Component decomposition of the TDSS for H33258 and molecular dynamics force field parameters for ground and excited state coumarin 102 are included. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- Faeder, J.; Ladanyi, B. M. *J. Phys. Chem. B* **2001**, *105*, 11148.
- Feng, X. B.; Thompson, W. H. *J. Phys. Chem. C* **2007**, *111*, 18060.
- Furse, K. E.; Corcelli, S. A. *J. Am. Chem. Soc.* **2008**, *130*, 13103.
- Golosov, A. A.; Karplus, M. *J. Phys. Chem. B* **2007**, *111*, 1482.
- Li, T. P.; Hassanali, A. A. P.; Kao, Y. T.; Zhong, D. P.; Singer, S. J. *J. Am. Chem. Soc.* **2007**, *129*, 3376.
- Nilsson, L.; Halle, B. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13867.
- Pal, S.; Maiti, P. K.; Bagchi, B.; Hynes, J. T. *J. Phys. Chem. B* **2006**, *110*, 26396.
- Sen, S.; Andreatta, D.; Ponomarev, S. Y.; Beveridge, D. L.; Berg, M. A. *J. Am. Chem. Soc.* **2009**, *131*, 1724.
- Thompson, W. H. *J. Chem. Phys.* **2004**, *120*, 8125.
- Furse, K. E.; Lindquist, B. A.; Corcelli, S. A. *J. Phys. Chem. B* **2008**, *112*, 3231.
- Lesch, H.; Schlichter, J.; Friedrich, J.; Vanderkooi, J. M. *Biophys. J.* **2004**, *86*, 467.
- Muino, P. L.; Callis, P. R. *J. Chem. Phys.* **1994**, *100*, 4093.
- Roux, B.; Simonson, T. *Biophys. Chem.* **1999**, *78*, 1.
- Sagui, C.; Darden, T. A. *Annu. Rev. Biophys. Biomol. Struct.* **1999**, *28*, 155.
- Tobias, D. J. *Curr. Opin. Struct. Biol.* **2001**, *11*, 253.
- Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089.
- Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577.
- Ewald, P. P. *Ann. Phys.* **1921**, *64*, 253.
- Onsager, L. *J. Am. Chem. Soc.* **1936**, *58*, 1486.
- Born, M. *Z. Phys.* **1920**, *1*, 45.
- Grossfield, A.; Sachs, J.; Woolf, T. B. *Proteins* **2000**, *41*, 211.
- Rokhlin, V. *J. Comput. Phys.* **1985**, *60*, 187.
- Ladd, A. J. C. *Mol. Phys.* **1977**, *33*, 1039.
- Ladd, A. J. C. *Mol. Phys.* **1978**, *36*, 463.
- Neumann, M. *Mol. Phys.* **1987**, *60*, 225.
- Neumann, M.; Steinhauser, O.; Pawley, G. S. *Mol. Phys.* **1984**, *52*, 97.
- Ladanyi, B. M.; Stratt, R. M. *J. Phys. Chem.* **1995**, *99*, 2502.
- Wolf, D.; Keblinski, P.; Phillpot, S. R.; Eggebrecht, J. *J. Chem. Phys.* **1999**, *110*, 8254.
- Fennell, C. J.; Gezelter, J. D. *J. Chem. Phys.* **2006**, *124*, 234104.
- Zahn, D.; Schilling, B.; Kast, S. M. *J. Phys. Chem. B* **2002**, *106*, 10725.
- Pal, S. K.; Zhao, L. A.; Zewail, A. H. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 8113.
- Andreatta, D.; Lustres, J. L. P.; Kovalenko, S. A.; Ernsting, N. P.; Murphy, C. J.; Coleman, R. S.; Berg, M. A. *J. Am. Chem. Soc.* **2005**, *127*, 7270.
- Andreatta, D.; Sen, S.; Lustres, J. L. P.; Kovalenko, S. A.; Ernsting, N. P.; Murphy, C. J.; Coleman, R. S.; Berg, M. A. *J. Am. Chem. Soc.* **2006**, *128*, 6885.
- Berg, M. A.; Coleman, R. S.; Murphy, C. J. *Phys. Chem. Chem. Phys.* **2008**, *10*, 1229.

- (35) Brauns, E. B.; Madaras, M. L.; Coleman, R. S.; Murphy, C. J.; Berg, M. A. *J. Am. Chem. Soc.* **1999**, *121*, 11644.
- (36) Brauns, E. B.; Madaras, M. L.; Coleman, R. S.; Murphy, C. J.; Berg, M. A. *Phys. Rev. Lett.* **2002**, *88*, 158101.
- (37) Brauns, E. B.; Murphy, C. J.; Berg, M. A. *J. Am. Chem. Soc.* **1998**, *120*, 2449.
- (38) Coleman, R. S.; Berg, M. A.; Murphy, C. J. *Tetrahedron* **2007**, *63*, 3450.
- (39) Gearheart, L. A.; Somoza, M. M.; Rivers, W. E.; Murphy, C. J.; Coleman, R. S.; Berg, M. A. *J. Am. Chem. Soc.* **2003**, *125*, 11812.
- (40) Sen, S.; Gearheart, L. A.; Rivers, E.; Liu, H.; Coleman, R. S.; Murphy, C. J.; Berg, M. A. *J. Phys. Chem. B* **2006**, *110*, 13248.
- (41) Sen, S.; Paraggio, N. A.; Gearheart, L. A.; Connor, E. E.; Issa, A.; Coleman, R. S.; Wilson, D. M.; Wyatt, M. D.; Berg, M. A. *Biophys. J.* **2005**, *89*, 4129.
- (42) Somoza, M. M.; Andreatta, D.; Murphy, C. J.; Coleman, R. S.; Berg, M. A. *Nucleic Acids Res.* **2004**, *32*, 2494.
- (43) Beck, D. A. C.; Armen, R. S.; Daggett, V. *Biochemistry* **2005**, *44*, 609.
- (44) Feller, S. E.; Pastor, R. W.; Rojnuckarin, A.; Bogusz, S.; Brooks, B. R. *J. Phys. Chem.* **1996**, *100*, 17011.
- (45) Hunenberger, P. H.; McCammon, J. A. *J. Chem. Phys.* **1999**, *110*, 1856.
- (46) Weber, W.; Hunenberger, P. H.; McCammon, J. A. *J. Phys. Chem. B* **2000**, *104*, 3668.
- (47) Lins, R. D.; Rothlisberger, U. *J. Chem. Theory Comput.* **2006**, *2*, 246.
- (48) Monticelli, L.; Simoes, C.; Belvisi, L.; Colombo, G. *J. Phys.: Condens. Matter* **2006**, *18*, S329.
- (49) Carter, E. A.; Hynes, J. T. *J. Chem. Phys.* **1991**, *94*, 5961.
- (50) Laird, B. B.; Thompson, W. H. *J. Chem. Phys.* **2007**, *126*, 211104.
- (51) Biswas, R.; Bagchi, B. *J. Phys. Chem. A* **1999**, *103*, 2495.
- (52) Friedman, H. L.; Raineri, F. O.; Hirata, F.; Perng, B. C. *J. Stat. Phys.* **1995**, *78*, 239.
- (53) Friedman, H. L.; Raineri, F. O.; Perng, B. C. *J. Mol. Liq.* **1995**, *7*, 65–6.
- (54) Ingrosso, F.; Ladanyi, B. M. *J. Phys. Chem. B* **2006**, *110*, 10120.
- (55) Ingrosso, F.; Ladanyi, B. M.; Mennucci, B.; Elola, M. D.; Tomasi, J. *J. Phys. Chem. B* **2005**, *109*, 3553.
- (56) Ingrosso, F.; Tani, A.; Tomasi, J. *J. Mol. Liq.* **2005**, *117*, 85.
- (57) Ishida, T.; Hirata, F.; Kato, S. *J. Chem. Phys.* **1999**, *110*, 11423.
- (58) Kumar, P. V.; Maroncelli, M. *J. Chem. Phys.* **1995**, *103*, 3038.
- (59) Ladanyi, B. M.; Klein, S. *J. Chem. Phys.* **1996**, *105*, 1552.
- (60) Ladanyi, B. M.; Maroncelli, M. *J. Chem. Phys.* **1998**, *109*, 3204.
- (61) Ladanyi, B. M.; Perng, B. C. *J. Phys. Chem. A* **2002**, *106*, 6922.
- (62) Ladanyi, B. M.; Stratt, R. M. *J. Phys. Chem.* **1996**, *100*, 1266.
- (63) Lobaugh, J.; Rossky, P. J. *J. Phys. Chem. A* **1999**, *103*, 9432.
- (64) Maroncelli, M. *J. Chem. Phys.* **1991**, *94*, 2084.
- (65) Nishiyama, K.; Yamaguchi, T.; Hirata, F.; Okada, T. *Pure Appl. Chem.* **2004**, *76*, 71.
- (66) Raineri, F. O.; Perng, B. C.; Friedman, H. L. *Chem. Phys.* **1994**, *183*, 187.
- (67) Roy, S.; Komath, S.; Bagchi, B. *J. Chem. Phys.* **1993**, *99*, 3139.
- (68) Sudholt, W.; Staib, A.; Sobolewski, A. L.; Domcke, W. *Phys. Chem. Chem. Phys.* **2000**, *2*, 4341.
- (69) Wang, J. M.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049.
- (70) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- (71) Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785.
- (72) Miehlich, B.; Savin, A.; Stoll, H.; Preuss, H. *Chem. Phys. Lett.* **1989**, *157*, 200.
- (73) Cave, R. J.; Castner, E. W. *J. Phys. Chem. A* **2002**, *106*, 12117.
- (74) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269.
- (75) Ravi, M.; Soujanya, T.; Samanta, A.; Radhakrishnan, T. P. *J. Chem. Soc., Faraday Trans.* **1995**, *91*, 2739.
- (76) Samanta, A.; Fessenden, R. W. *J. Phys. Chem. A* **2000**, *104*, 8577.
- (77) Ladinig, M.; Leupin, W.; Meuwly, M.; Respondek, M.; Wirz, J.; Zoete, V. *Helv. Chim. Acta* **2005**, *88*, 53.
- (78) Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; Schafmeister, C.; Ross, W. S.; Kollman, P. A. *AMBER 9*; University of California: San Francisco, CA; 2006.
- (79) Berendsen, H. J. C.; Grigera, J. R.; Straatsma, T. P. *J. Phys. Chem.* **1987**, *91*, 6269.
- (80) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327.
- (81) Stratt, R. M.; Maroncelli, M. *J. Phys. Chem.* **1996**, *100*, 12981.
- (82) Lustres, J. L. N.; Kovalenko, S. A.; Mosquera, M.; Senyushkina, T.; Flasche, W.; Ernsting, N. P. *Angew. Chem., Int. Ed.* **2005**, *44*, 5635.
- (83) Karunakaran, V.; Pfaffe, M.; Ioffe, I.; Senyushkina, T.; Kovalenko, S. A.; Mahrwald, R.; Fartzdinov, V.; Sklenar, H.; Ernsting, N. P. *J. Phys. Chem. A* **2008**, *112*, 4294.
- (84) Jimenez, R.; Fleming, G. R.; Kumar, P. V.; Maroncelli, M. *Nature* **1994**, *369*, 471.

Monte Carlo Sampling with Hierarchical Move Sets: POSH Monte Carlo

Jerome Nilmeier* and Matthew P. Jacobson

*Graduate Group in Biophysics, University of California, San Francisco,
California 94158*

Received November 30, 2008

Abstract: We present a new Monte Carlo method for sampling rugged energy landscapes that allows for efficient transitions across sparsely distributed local basins. The trial move consists of two steps. The first step is a large initial trial move, and the second step is a Monte Carlo trajectory generated using smaller trial moves. To maintain detailed balance, a reverse transition probability is estimated along a path that differs from the forward path. Since the forward and reverse transitions are different, we name the algorithm POSH (port out, starboard home) Monte Carlo. The process obeys detailed balance to the extent that the transition probabilities are correctly estimated. There is an optimal range of performance for a given energy landscape, which depends on how sparsely the low energy states of the system are distributed. For simple model systems, adequate precision is obtained over a large range of inner steps settings. Side chain sampling of residues in the binding region of progesterone antibody 1dba are studied, and show that significant improvement over a comparable standard protocol can be obtained using POSH sampling. To compare with experimental data, the phosphopeptide Ace-Gly-Ser-pSer-Ser-Nma is also studied, and the resulting NMR observables compare well with experiment. For the biomolecular systems studied, we show that POSH sampling generates precise distributions using the number of inner steps set up to 20.

Introduction

The Metropolis algorithm¹ has been in use for over 50 years, with generalizations of the idea applied to fields far beyond the field of molecular modeling, for which it was initially developed. For systems with densely packed geometries, the generation of good trial moves can be a substantial challenge. A variety of methods in the field of liquid simulation have been developed to address this type of problem.^{2–4} For polymeric systems, chain growth methods,^{5,6} pivot moves,⁷ and other methods have been developed using the idea of biased sampling in polymeric^{8–10} and continuum contexts.^{11,12}

Many sampling algorithms have been designed expressly for crossing large energetic barriers and obeying, or nearly obeying, detailed balance. Most of these involve generating a Markov chain of states (a walker) in an expanded state space,¹³ which can be accepted to the configuration space of interest based on a modified acceptance probability.^{14–18}

The most widely adopted of these types is the replica exchange¹⁹ method,²⁰ and other closely related methods,^{21–24} which are particularly well suited to parallel processing.

The landscapes of biomolecular systems are rugged, which has motivated development of methods designed to locate global minima. For biomolecular systems, molecular dynamics is widely used, but it has been shown that all atom MC approaches can provide significant improvements in sampling.²⁵ Methods such as Monte Carlo Minimization²⁶ have emerged as a practical method for sampling landscapes that have multiple, sparsely distributed minima. This produces an approximately correct distribution while sampling a broader range of configurations. The approximation results in an underestimate of the entropy of local basins. More rigorous estimates of basin entropy can be made, as is done with the Mining Minima approach.^{27,28} In many cases, particularly in difficult biomolecular optimization problems, such as protein folding,^{29,30} drug design, and homology modeling,^{31–34} the use of minimization and other optimiza-

* To whom correspondence should be addressed. E-mail: jerome.nilmeier@ucsf.edu.

tion approaches have become the mainstay of the field, with entropic considerations added as a secondary effect, if at all.

In this work, we propose a simple, general method for sampling rugged landscapes that obeys detailed balance to the extent that the transition probabilities are correctly estimated. The estimate of the transition probabilities is discussed in the theory section. The basic idea is to generate a large initial trial perturbation, followed by a series of small Monte Carlo moves to anneal the initial trial move to a lower energy, and accept the resulting trial move with a modified acceptance probability.

Motivation

Partitioning of Configuration Space. A Monte Carlo sampling strategy seeks to evaluate integrals of the type

$$\begin{aligned} \langle O \rangle &= \frac{1}{Z} \int d\mathbf{q} O(\mathbf{q}) \exp(-\beta U(\mathbf{q})) \\ Z &= \int d\mathbf{q} \exp(-\beta U(\mathbf{q})) \end{aligned} \quad (1)$$

where \mathbf{q} is the set of all coordinates of the system of interest, $\beta = (k_B T)^{-1}$ is the inverse temperature, and $U(\mathbf{q})$ is the potential energy. The observed quantity $\langle O \rangle$ is averaged over many instances of $O(\mathbf{q})$. Z is the normalization constant or configuration integral.

The Boltzmann factor in the integrand is the unnormalized probability distribution:

$$p(\mathbf{q}) = \exp(-\beta U(\mathbf{q})) \quad (2)$$

Often, a natural partition of the entire space becomes convenient. A common example is the Born–Oppenheimer approximation, where the nuclear degrees of freedom are considered to be uncoupled or adiabatic relative to the electronic degrees of freedom.⁶ This type of approximation also appears in the formulation of implicit solvation models, where the solvent degrees of freedom are integrated out, and an approximate model for the interaction between a macromolecule and the solvent is introduced.³⁵ Propagation along adiabatic degrees of freedom has been introduced in both Monte Carlo^{11,36,37} and dynamical^{36–38} contexts. For the applications motivating this work, the coordinate decomposition is between protein backbone and side chain coordinates.

The motivation for sampling separate subspaces is often guided by the definition of a configuration space partitioning where the covariant fluctuations between the partitioned subspaces are small. This can often be justified by a dynamical argument, as is the case for the examples given above. The practical motivation for decomposition of subspaces is computational efficiency. In proteins, for example, different geometric algorithms are appropriate for sampling backbone^{39,40} and side chain degrees of freedom,⁴¹ and the challenge lies in combining these trial moves in a way that preserves ergodicity, while generating a high acceptance ratio and, of course, the expected distribution of states.

Figure 1 shows a schematic to motivate the development of a sampling protocol. The complete configurational space is partitioned into torsional coordinates of the backbone (ϕ) and torsional coordinates of the side chains (χ). Consider a trial move that consists of randomly selecting a subspace to

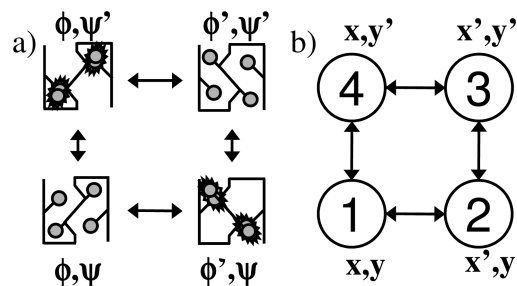


Figure 1. (a) Schematic of the multiple subspace sampling problem for proteins. Backbone and side chain coordinates are ϕ and χ , respectively. (b) Generalization of the 2 subspace sampling problem. \mathbf{x} and \mathbf{y} represent subspaces to be sampled.

perturb (ϕ or χ), and accepting with the Metropolis criterion. A perturbation in backbone space may generate a configuration with the side chains in a high energy state. Likewise, a perturbation of side chains only may also lead to a high energy state. Although a series of samples from the trial state could generate a lower energy state, the initial move would be rejected outright with a standard scheme. If both degrees of freedom are perturbed simultaneously, however, the likelihood of generating a reasonable trial can become vanishingly small. A system that remains in a macrostate for long times relative to local correlations is often referred to as quasi-ergodic,¹⁵ frustrated/glassy,⁴² or kinetically trapped.

For the remainder of the work, the *original* coordinate state will be labeled as state 1, the *initial trial state* will be state 2, and the *final trial state* will be labeled state 3. The *reflected trial state*, labeled as state 4, will be defined shortly. These states are illustrated in Figure 2. In the cases shown in Figure 1, the most intuitive solution is to generate a series of Monte Carlo steps in the alternate subspace to locate a low energy coordinate and then accept the final trial coordinate with some reasonable probability. In this sense, the transition from the initial trial to the final trial can be thought of as an annealing step.

Theory

Detailed Balance. To determine the proper acceptance criterion, the condition of detailed balance is:

$$p_i T_{ij} = p_j T_{ji} \quad (3)$$

where i and j are two arbitrary coordinate states with probabilities $p_i = p(\mathbf{q}_i)$ and $p_j = p(\mathbf{q}_j)$, as given by eq 2, respectively. The transition probability T_{ij} is the probability of transitioning to coordinate state j from state i , and T_{ji} is the reverse transition probability. The condition of detailed balance will be applied to the states 1 and 3, as shown in Figure 2

$$p_1 T_{13} = p_3 T_{31} \quad (4)$$

where the forward transition consists of a trial move followed by a chain of moves. In the cases described by Figure 2a and b, for example, the initial trial could be also notated as $p_2 = P(\mathbf{x}', \mathbf{y})$, where \mathbf{x}' is a trial move, and so on, as labeled.

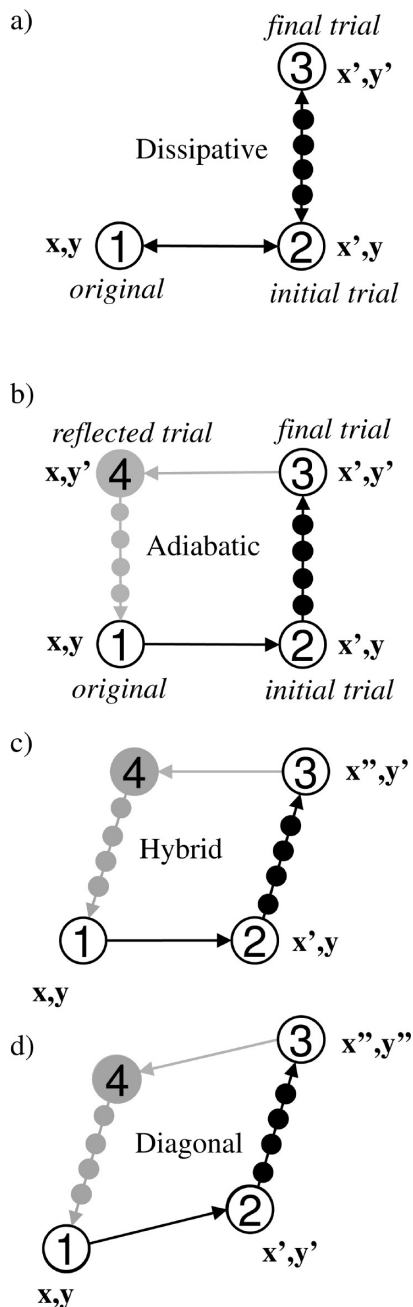


Figure 2. Topological representation of detailed balance condition (eq 4): (a) dissipative pathway (eq 5) and (b–d) push pathways (eq 12), (b) adiabatic pathway, (c) hybrid pathway, and (d) diagonal pathway. Calculation of coordinate 4 is described in the text and in Figure 3. Coordinate labels correspond to state numbers.

The condition of detailed balance is met only if we choose to enforce the flow rates between states 1 and 3. If all states were accounted for (1–4), then only the condition of balance would be satisfied.⁴³ In either case, however, the satisfaction of eq 4 will ensure a proper distribution of states.

Hierarchical Perturbations. The general scheme is a hierarchical decomposition of move sets. The initial perturbation is designed to be a large move which will cover large regions of configurational space, while the series of trial moves in the “inner loop” are much smaller and designed to be an annealing move. The initial trial move, labeled $\xi^{(1,2)}$, is, for all cases here, a vector of uniform variates over some

domain $[-\mathbf{d}^{(1,2)}/2, \mathbf{d}^{(1,2)}/2]$, where $\mathbf{d}^{(1,2)}$ is a vector of the same dimension as the complete space (with zero entries for the degrees of freedom that are not sampled). The series of Monte Carlo steps in the inner loop use a different perturbation type, $\xi_n^{(2,3)}$, where n is the inner step number, and the domain is $[-\mathbf{d}^{(2,3)}/2, \mathbf{d}^{(2,3)}/2]$.

Figure 2 shows schematically the ways that the perturbations can differ. The first pathway, labeled as the dissipative pathway, is used frequently in nonequilibrium studies. For the adiabatic pathway, the initial perturbation is only within one subspace, \mathbf{x} , while the annealing steps are within the \mathbf{y} subspace only. For the hybrid pathway, the initial perturbations are in \mathbf{x} , while the remaining perturbations are in either \mathbf{x} or \mathbf{y} (or both). Finally, the diagonal path allows for perturbations in both \mathbf{x} and \mathbf{y} in both the initial and final step, with the differences being in the magnitude of the perturbations.

The essential feature of the algorithm presented is that the perturbation domains $\mathbf{d}^{(1,2)}$ and $\mathbf{d}^{(2,3)}$ perturbations differ in some hierarchical way. Typically, the initial perturbations will be larger, such that local basins can be traversed. The annealing step uses smaller perturbations to search for nearby low energy states from the initial trial. As long as the total space is covered by the combination of perturbations, meaning that there are no zero values in the vector sum $\mathbf{d}^{(1,2)} + \mathbf{d}^{(2,3)}$, the complete sampling of space is possible.

This approach is not, in and of itself, guaranteed to solve the quasi-ergodic problem, however. The connectivity between states that is generated by perturbation domain $\mathbf{d}^{(1,2)}$ can still limit the accessibility of alternative macrostates. The use of the annealing steps, however, permits larger move sets in the initial trial step that might not otherwise be practical. Table 1 describes the types of move sets used for all simulations generated for the present work.

Acceptance Criteria. The forward transition can be defined as a combination of moves described diagrammatically in Figure 2. Consider first the forward transition probability, which for all cases follows the pathway (1 → 2 → 3), which is a combination of the (1 → 2) transition and the (2 → 3) transition. The difference in the nature of the reverse pathway determines the two acceptance criteria presented.

Dissipative Pathway and Acceptance Criterion. The standard approach from this point would be to require that the reverse transition be along the pathway (3 → 2 → 1). See Figure 2a. For this case, the condition for detailed balance is

$$p_1 \alpha_{12} \alpha_{23} \text{acc}_{13} = p_3 \alpha_{12} \alpha_{32} \text{acc}_{31} \quad (5)$$

where α_{ij} and acc_{ij} are the selection and acceptance probabilities, respectively, of state j from state i . This particular condition of detailed balance has been invoked in the study of nonequilibrium systems^{43,44} and is labeled here as the dissipative pathway. The trial coordinate \mathbf{q}_2 is generated as

$$\mathbf{q}_2 = \mathbf{q}_1 + \xi^{(1,2)} \quad (6)$$

where $\xi^{(1,2)}$ is a uniform deviate vector that perturbs along the (1 → 2) portion of the pathway. Figure 3a shows the

Table 1. Simulation Settings for Simulations of Model Systems^a

figure	B	$\mathbf{d}_x^{(1,2)}$	$\mathbf{d}_y^{(1,2)}$	$\mathbf{d}_x^{(2,3)}$	$\mathbf{d}_y^{(2,3)}$	δb	path type	reverse pathway
5	45	22.5	22.5	9	9	1.0	diagonal	true
6	60	60	60	12	12	0.25	diagonal	true
11	45	22.5	0	0	9	0.9	adiabatic	true
11	45	22.5	0	9	9	0.9	hybrid	true
11	45	22.5	22.5	9	9	0.9	diagonal	true
12	45	22.5	22.5	9	9	0.9	diagonal	both

^a Parameters of landscapes as described in Appendix A. B represents the lower and upper bound of the landscape in both the domains of both x and y . The values in columns 3 and 4 are the lower and upper bounds of the initial ($1 \rightarrow 2$) perturbation for x and y , respectively, and columns 5 and 6 are the lower and upper bounds for the ($2 \rightarrow 3$) perturbation for x and y , respectively. δb is the square bin width of the 2D histograms collected.

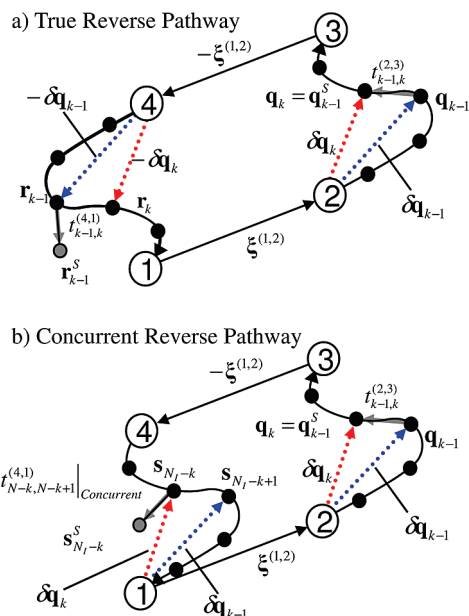


Figure 3. Notation and indexing for transition matrix construction. Coordinate states 1–4 are described in the text. The perturbation from 1 to 2 is given by the vector $\xi^{(1,2)}$, which also connects states 3 and 4. The forward trajectory along the ($2 \rightarrow 3$) pathway is connected by a line with black dots, which represent coordinates along the annealed path. A set of trials at step ($k-1 \rightarrow k$) is shown for reference. For clarity, the case of an accepted trial move in both forward and reverse transitions are shown. For the forward trajectory, the accepted move is a black dot connected with a gray arrow, given by eqs 17 or 19. The gray arrow represents the transition for which a probability is computed. The red and blue arrows show the difference vectors for which the reverse trajectory is constructed (eq 14). The reverse transition probability connects the $k-1$ reverse coordinate to an accepted trial coordinate, which is not necessarily the same as the next step in the reverse pathway. The only difference between a and b is in the way that the reverse coordinate is constructed using the difference vectors (shown in red and blue): (a) true reverse pathway (eq 22) and (b) concurrent reverse pathway (eq 24).

forward path construction in detail. The final trial coordinate \mathbf{q}_3 is generated using a series of Monte Carlo transitions.

The selection probabilities along the ($2 \rightarrow 3$) and ($3 \rightarrow 2$) are given by the transition probabilities⁴⁵

$$\begin{aligned} \alpha_{23} &= T_{23}^{(N_i)} \\ \alpha_{32} &= T_{32}^{(N_i)} \end{aligned} \quad (7)$$

where $T_{23}^{(N_i)}$ and $T_{32}^{(N_i)}$ are the transition probabilities along the respective Markov chains of length N_i . Returning to eq 5, we can see that the probability of selecting state 3 from state 1 is the probability of selecting the trial move ($1 \rightarrow 2$) and selecting the final trial ($2 \rightarrow 3$). The joint probability of this selection process is represented by the product of the selection probabilities. The final trial ($2 \rightarrow 3$) selection probability is determined by the transition probability of the inner loop sampling procedure. Equation 7 simply states that the probability of selecting state j from state i is the transition probability of the Markov Chain. The transition probability of a multistep stochastic walk is described by the Chapman–Kolmogorov equation

$$T_{ij}^{(N_i)} = T^{(N_i)}(\mathbf{q}_j|\mathbf{q}_i) = \int d\mathbf{q}_1 \dots d\mathbf{q}_{N_i-1} \prod_{k=1}^{N_i} t^{(ij)}(\mathbf{q}_{k-1}|\mathbf{q}_k) \quad (8)$$

where $k=0$ corresponds to initial state i and $k=N_i$ corresponds to state j , and $t^{(ij)}(\mathbf{q}_k|\mathbf{q}_{k-1}) = t_{k-1,k}^{(ij)}$ is the transition probability from state $k-1$ to state k (at step k)

$$T_{ij}^{(N_i)} = \prod_{k=1}^{N_i} t_{k-1,k}^{(ij)} \quad (9)$$

Equations 8 and 9 form an exact description of the transition probability, which incorporates all possible paths connecting states i and j , if the transition elements $t_{k-1,k}^{(ij)}$ are known rigorously.

If each inner step obeys detailed balance, the resulting detailed balance condition is satisfied⁴⁶

$$p_2 T_{23}^{(N_i)} = p_3 T_{32}^{(N_i)} \quad (10)$$

Solving eq 27 gives

$$\frac{\text{acc}_{13}}{\text{acc}_{31}} = \frac{p_3 T_{32}^{(N_i)}}{p_1 T_{23}^{(N_i)}} = \frac{p_2}{p_1} \quad (11)$$

This criterion, while not originally derived as a Monte Carlo acceptance criterion, is an important criterion because it provides a foundation for thinking of these systems in broader terms. This criterion can be thought of as the work required to place the system in a trial state. This represents a rigorously correct acceptance criterion that does not require the estimation of a transition probability, which can give insight into limiting behavior.

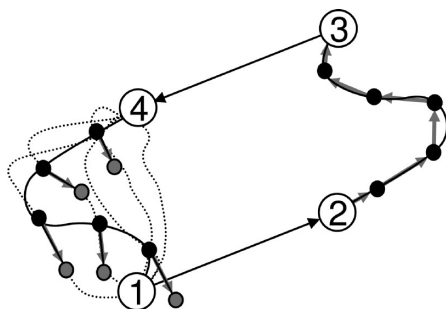


Figure 4. Illustration of forward and reverse transition probability calculations. Dotted lines represent putative trajectories that exhibit similar transitions in various stages of the trajectory.

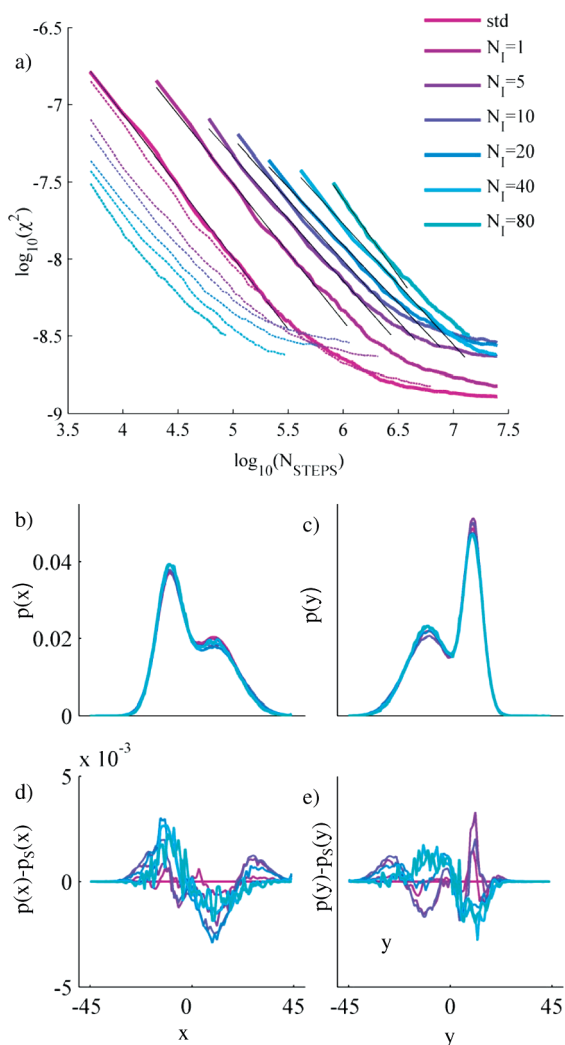


Figure 5. Convergence behavior of simple model system. Model system is landscape D. (a) log–log plot of convergence rates. Solid lines are shown for $N_{\text{STEPS}} = N_0(2N_i + 1)$, where N_0 is the number of outer steps, and dotted lines are plotted against $N_{\text{STEPS}} = N_0$. (b) PMFs along x coordinate in normalized probability units (c) PMFS along y (d, e) residual probability from standard simulation PMFs.

posh Pathway and Acceptance Criterion. Consider the possibility of the reverse pathway passing through a different region of coordinate space. This type of consideration has

been made in studying alternative transitions of trajectories in phase space.⁴⁷ To emphasize this difference in forward and reverse paths, we describe this as the posh pathway (port out, starboard home). Here, state 4 consists of the reverse trial perturbation, followed by a trajectory, which arrives at state 1. This can be described as

$$p_1\alpha_{12}\alpha_{23}\text{acc}_{13} = p_3\alpha_{34}\alpha_{41}\text{acc}_{31} \quad (12)$$

This condition for detailed balance as shown in Figure 2b was originally stated by Siepmann^{11,12} in the development of the adiabatic nuclear and electronic sampling MC method. In his derivation, the degrees of freedom for the initial perturbation were orthogonal to the annealed degrees of freedom, under the adiabatic assumption. Defining the topology of the states is sufficient to derive a slightly more general form, with the assignment of coordinate states added only for clarity. The 3 cases of interest are described in Figure 2b–d.

Coordinate state 4, the reflected trial coordinate \mathbf{q}_4 , can be defined in terms of the final trial coordinate and the uniform deviate $\xi^{(1,2)}$ in eq 6 used to generate the trial coordinate

$$\begin{aligned} \mathbf{q}_4 &= \mathbf{q}_3 - \xi^{(1,2)} \\ \mathbf{q}_4 &= \mathbf{q}_1 + \delta\mathbf{q}_{N_i} \end{aligned} \quad (13)$$

where

$$\delta\mathbf{q}_{N_i} = \mathbf{q}_3 - \mathbf{q}_2 \quad (14)$$

is the change in position from the trial position to the final trial position after a series of inner steps N_i .

In principle, the coordinate state \mathbf{q}_4 need not be defined in terms of the forward random deviate to preserve the property $\alpha_{12} = \alpha_{34}$. The choice of the reflected trial coordinate given by eq 13 has the convenient property of allowing the transition pathway ($4 \rightarrow 1$) to be defined using the information from the forward trajectory, forming a closed loop. The ($2 \rightarrow 3$) and ($4 \rightarrow 1$) selection probabilities are given by the following transition probabilities⁴⁶

$$\begin{aligned} \alpha_{23} &= T_{23}^{(N_i)} \\ \alpha_{41} &= T_{41}^{(N_i)} \end{aligned} \quad (15)$$

where $T_{23}^{N_i}$ and $T_{41}^{N_i}$ are the transition probabilities along the respective Markov chains of length N_i . Solving for the ratio of acceptance probabilities gives

$$\frac{\text{acc}_{13}}{\text{acc}_{31}} = \frac{p_3 T_{41}^{(N_i)}}{p_1 T_{23}^{(N_i)}} \quad (16)$$

Equation 16 obeys detailed balance to the extent that the transition probabilities are correctly estimated. The estimates of the transition probabilities are described in the next sections.

Forward Transition Probability. The forward transition probability is estimated using the record of the selection and acceptance probabilities of the single trajectory generated. While this is only an estimate of eq 8, this type of approach

is commonly used in other biased sampling methods^{6,13,14,47} in varying contexts. For the purposes of this work, it is assumed that the selection probability is uniform for all trial moves, so that only acceptance probabilities need to be computed. The acceptance probabilities presented are either the Barker or the Metropolis–Hastings acceptance probability.^{8,46} For the Barker acceptance probability, uniform variate is generated over the domain $[0, p(\mathbf{q}_{k-1}) + p(\mathbf{q}'_{k-1})]$. If the random number is greater than or equal to $p(\mathbf{q}'_{k-1})$, the trial move is accepted, and $\mathbf{q}_{k-1}^S = \mathbf{q}'_{k-1}$. Otherwise, the trial move is rejected, and $\mathbf{q}_{k-1}^S = \mathbf{q}_{k-1}$. If a trajectory in the inner loop is generated using the Barker acceptance criterion, the transition probability at the k th inner step is

$$t_{k-1,k}^{(2,3)} = \frac{p(\mathbf{q}_k = \mathbf{q}_{k-1}^S)}{p(\mathbf{q}_{k-1}) + p(\mathbf{q}'_{k-1})} \quad (17)$$

where the trial coordinate \mathbf{q}'_{k-1} is generated using a uniform deviate vector

$$\mathbf{q}'_{k-1} = \mathbf{q}_{k-1} + \xi_k^{(2,3)} \quad (18)$$

The selected coordinate \mathbf{q}_{k-1}^S is the coordinate resulting from the application of the acceptance probability (shown here as the Barker criterion), which is \mathbf{q}'_{k-1} if the trial is accepted and \mathbf{q}_{k-1} if the trial is rejected.

Rewriting the transitions of eq 17 with the Metropolis transition probability gives:

$$(t_{k-1,k}^{(2,3)})_{\text{Metropolis}} = \delta(\mathbf{q}_{k-1}^S - \mathbf{q}'_{k-1}) \text{acc}_{k-1,k}^{(2,3)} + \delta(\mathbf{q}_{k-1}^S - \mathbf{q}_{k-1}) (1 - \text{acc}_{k-1,k}^{(2,3)}) \quad (19)$$

Dellago uses a similar relation to define the Metropolis action.⁴⁸ The usual Metropolis criterion is applied to obtain the selected coordinate \mathbf{q}_{k-1}^S

$$(\text{acc}_{k-1,k}^{(2,3)})_{\text{Metropolis}} = \min(1, p(\mathbf{q}'_{k-1})/p(\mathbf{q}_{k-1})) \quad (20)$$

The notation and generation of the trial coordinate of eq 19 are identical to that of eq 17. In principle, any record of transitions can be used in lieu of either of the two expressions presented. In practice, however, the relations which obey detailed balance have given good results for the model systems studied. Using Metropolis transitions is the generally preferable approach for most types of simulations,⁵⁰ and this is true for the present case as well. For the remainder of the paper, however, the transition elements will be described using the Barker acceptance probability, since the notation is easier to read.

Reverse Transition Estimates. It should be noted that the estimate of the $(4 \rightarrow 1)$ transition probability represents a challenging class of problems whereby the end points are known, and the calculation of all paths connecting them as described by eq 8 cannot be enumerated practically. In general, the estimate of this probability is accomplished through importance sampling. Techniques such as Transition Path Sampling, have gained wide use in generating such estimates,^{49,51,52} whereby macrostate end points are defined, rather than fixed coordinate states. This is a conceptually helpful way to consider the transition probability estimates,

especially since the forward $(2 \rightarrow 3)$ trajectory can be thought of as an importance-sampled transition path.

Reverse Transition Pathway (True). Figure 3a describes the reverse transition path construction. Since it most closely mirrors the forward path, we label this as the true reverse transition pathway. To construct the reverse pathway, we first keep track of the displacement from initial trial state at step k

$$\delta \mathbf{q}_k = \mathbf{q}_k - \mathbf{q}_2 \quad (21)$$

The reverse path is defined using the displacements from the forward path. The true reverse transition pathway can be estimated using the forward trajectory information in the following way:

$$\mathbf{r}_k = \mathbf{q}_4 - \delta \mathbf{q}_k \quad (22)$$

where \mathbf{r}_k is the k th coordinate state in the reverse pathway as constructed in Figure 3a. The transition probability from state $k - 1$ to k is recorded as

$$t_{k-1,k}^{(4,1)} = \frac{p(\mathbf{r}_{k-1}^S)}{p(\mathbf{r}_{k-1}) + p(\mathbf{r}'_{k-1})} \quad (23)$$

where $\mathbf{r}'_{k-1} = \mathbf{r}_{k-1} + \xi_k^{(4,1)}$ is constructed using the same perturbation strategy (using domain $\mathbf{d}^{(2,3)}$) as the forward case. The selection of coordinate \mathbf{r}_{k-1}^S follows the same procedure as in the forward pathway. It is important to notice that the selected coordinate $\mathbf{r}_{k-1}^S \neq \mathbf{r}_k$, since $p(\mathbf{r}_k)$ can easily become vanishingly small relative to $p(\mathbf{r}_{k-1})$ in the reverse trajectory for complex landscapes, whereas $p(\mathbf{r}_{k-1}^S)$ is selected according to its probability weight. The coordinate $\mathbf{r}_{k-1} = \mathbf{q}_4 - \delta \mathbf{q}_{k-1}$ forms the anchor point at each step along the reverse pathway, from which a trial coordinate is generated.

One practical consideration when using eqs 21–23 to generate the reverse trajectory is that it requires the storage of the complete forward trajectory $\delta \mathbf{q} = \{\delta \mathbf{q}_1, \delta \mathbf{q}_2, \dots, \delta \mathbf{q}_{N_1}\}$ prior to generating the reverse trajectory. Since these trajectories are generated using random deviates, this can be accomplished by maintaining a list of the random seeds, rather than an exhaustive storage of coordinate states. Even with this approach, however, it is can be cumbersome to reconstruct the entire reverse trajectory only after the entire forward trajectory has completed.

Concurrent Reverse Transition Pathway. To simplify the storage requirements, an alternative path for the reverse coordinate as a set of vectors $\mathbf{s} = \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_{N_1}\}$ can be defined

$$\mathbf{s}_{N_1-k} = \mathbf{q}_1 + \delta \mathbf{q}_k \quad (24)$$

which provides the same overall connectivity between state 4 and state 1 as the set of vectors \mathbf{r} , given by eq 22 (see also Figure 3b). The difference is that the reverse pathway can be generated concurrently with the forward trajectory in this case. Using this alternative definition of the reverse coordinate path, the following transition can be defined:

$$t_{N_1-k, N_1-k+1}^{(4,1)} \Big|_{\text{Concurrent}} = \frac{p(\mathbf{s}_{N_1-k}^S)}{p(\mathbf{s}_{N_1-k}) + p(\mathbf{s}'_{N_1-k})} \quad (25)$$

This pathway can be generated as the forward trajectory is being generated (since it does not require knowledge of the final trial state). The storage requirements are much smaller for this pathway, and it is slightly easier to implement. It also appears to provide slightly better precision of sampling. A discussion of the errors introduced by using the concurrent versus the reverse pathway is in Appendix B.

Qualitative Justification for the Reverse Pathway Estimation. The primary motivation for using accepted trial moves (using either method) at each step in the reverse pathway is to maintain numerical stability. Since the reverse pathway is constructed in a region of space that has not been located using importance sampling, as is the case with the forward pathway, reconstructing the path exactly will generate vanishingly small probabilities for even the simplest of landscapes, such as those studied for this work. The fact that the states are no longer connected contiguously may in fact improve the estimate, as a collection of transitions along the reverse pathway is estimating an ensemble of reverse pathways (see Figure 4). In fact, the key challenge to improving this sampling strategy is an understanding of how to efficiently and accurately estimate these transition probabilities.

Error Metric. As a general measure of the quality of the sampled distribution versus the true distribution, we can define the following ergodicity metric which is commonly used in assessing sampling quality:^{14,52,53}

$$\chi^2 = \frac{1}{N_B} \sum_{i=1}^{\sqrt{N_B}} \sum_{j=1}^{\sqrt{N_B}} (G(\mathbf{x}_i, \mathbf{y}_j) - H(\mathbf{x}_i, \mathbf{y}_j))^2 \quad (26)$$

where χ^2 is the mean squared error (MSE) over the course of the entire simulation, N_B is the number of bins, and $G(\mathbf{x}_i, \mathbf{y}_j)$ is the normalized distribution as described in Appendix A. $H(\mathbf{x}_i, \mathbf{y}_j)$ is the normalized histogram at the square bin centered about $(\mathbf{x}_i, \mathbf{y}_j)$. The length and width of the 2D histogram bin are $\delta b \times \delta b$.

Results and Discussion

We first explore the precision and accuracy of posh sampling protocols using simple two-dimensional model systems. We next present two initial molecular applications: several side chains in a flexible antibody binding sites and a phosphorylated tetrapeptide.

Precision of 2D Model System. A variety of parameter settings are studied in detail, and presented in Appendix B. Here, we study a single landscape (landscape D of Table 6 and Figure 10). This landscape was chosen as a metric for precision, as it has a saddle point with interesting asymmetry that can often reveal systematic biases in sampling. It is also a landscape that can be easily sampled with a standard method, providing a control.

Figure 5 shows the convergence behavior of χ^2 versus the number of steps on a log–log plot. Parameter settings for this system appear in Table 1. The dotted lines are plotted

against the number of outer steps (N_O) required to generate a desired χ^2 value, while the solid lines are plotted against total the number of steps, $2N_O(N_I + 1)$, which is more representative of the rate of convergence, as it accounts for the total number of energy evaluations. For the standard simulation, the total number of energy evaluations was 25M, and the remaining simulations were scaled to be of comparable length. The first thing to notice is that the dotted lines appear to converge faster than the standard simulation but in fact do not converge faster when accounting for the total number of energy evaluations. This is not surprising for this particular model system. A more careful discussion of regimes efficiency in sampling is discussed in the following section. The next thing to notice is that the (log–log) convergence rates are initially roughly linear with the same slope for all simulations. The standard simulation, which should rigorously reproduce the distribution, deviates from a steady rate of convergence at around $N_O = 10^5$, leveling off at roughly 1.3×10^{-9} , as it appears to be reaching a limit of precision for this metric.

The remaining simulations to not converge as rapidly, but do appear to converge to reasonable values of χ^2 , with a maximum value of 2.7×10^{-9} . The $N_I = 1$ simulation appears to be approaching the limit achieved by the standard simulation, while values of $N_I = 5, 10$, and 20 appear to converge to a higher value, suggesting a slight bias in the sampling for these regions. Values of $N_I = 40$ and 80 appear to be converging to a lower value. A plot of the pmfs of landscape D are given in Figures 5b and 5c. Close inspection reveals some systematic bias in the sampling that is consistent with the values reported in Figure 5a. The errors that are apparent appear to be within a reasonable margin, however, as the residuals here reach a maximum value of 3×10^{-3} . The interesting change in behavior of the precision as the number of inner steps increases is due to a simple effect. The initial bias introduced comes from the overestimate of the value of the reverse transition probability, which results in distributions that have slightly higher entropy than the true distribution. (This effect is seen in the biomolecular systems as well, and is discussed in later sections.) As the number of inner steps become very large, such that each trial move is essentially fully equilibrated, the anomalously high reverse transition probability is offset by the fact that the trial move is annealed to an equilibrium state, and thus will not have an artificially high entropy. While the error is detectable here, it appears to be sufficiently small to assert that this approach obeys detailed balance.

Efficiency Considerations. To compare efficiency of different sampling methods, we compute the MSE generated using the same number of energy evaluations throughout the entire trajectory, including those energy evaluations in the inner loop. For an inner loop of length N_I , the number of energy evaluations required is $2(N_I + 1)$ per outer step. We can define a simple improvement metric to compare the efficiency of posh MC to a “standard” MC trajectory (no inner steps) as

$$I = \frac{\chi_s^2}{\chi^2(N_I)} \quad (27)$$

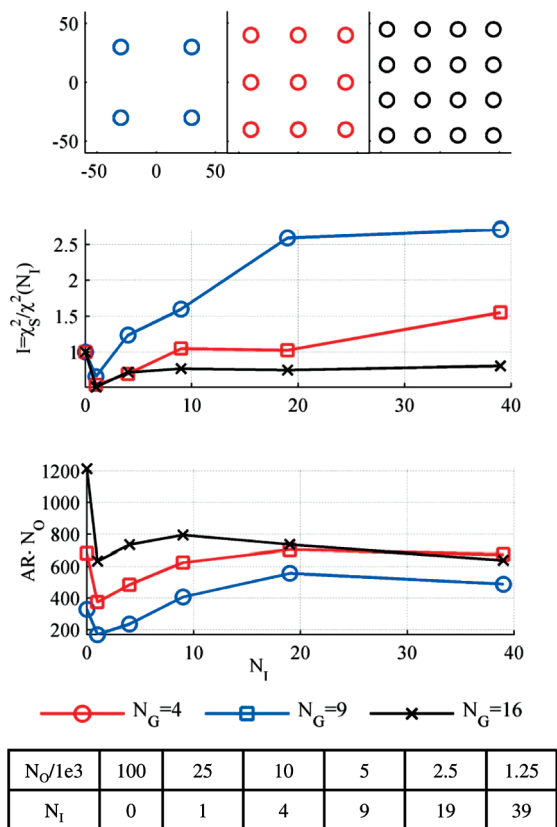


Figure 6. Efficiency of posh sampling depends on the sparsity of minima. (a) Schematic of landscapes sampled. Colors correspond to legend. Each circle represents a single Gaussian basin (Table 1, landscape G). (b) MSE versus number of inner steps. (c) Number of steps accepted. The number of outer steps in each case is adjusted such that the total number of energy evaluations is the same (100k) for each setting. See Table 6 for simulation settings.

where the MSE of a reference simulation, χ_s^2 , is computed using the same number of total energy evaluations as the posh MC simulation. In the posh MC simulations, the number of outer steps is thus $N_O = 2N_{STD}/(N_I + 1)$, where N_{STD} is the number of energy evaluations in a standard Monte Carlo simulation. Thus, while a larger number of inner steps will almost always improve the acceptance ratio, it will not always improve the efficiency of sampling, because of the cost of generating the trial move.

As a first evaluation of efficiency, we use very simple energy landscapes, represented schematically in Figure 8a. The 3 landscapes each have evenly distributed arrangements of equivalent basins in the same domain. We denote the number of basins as N_G . The basins are identical, each with parameters given as landscape G in Table 6.

For $N_G = 4$ (the energy landscape least densely populated with energy minima), the improvement in sampling efficiency increases with increasing number of inner steps, N_I . For $N_I = 4$, the posh sampling is actually less efficient than that of the standard simulation. There are simply not enough inner step moves to reliably locate nearby basins because the large initial trial move in the sparse space will rarely land in a favorable region of space. At $N_I = 9$, however, the annealing steps begin to locate basins and a noticeable improvement

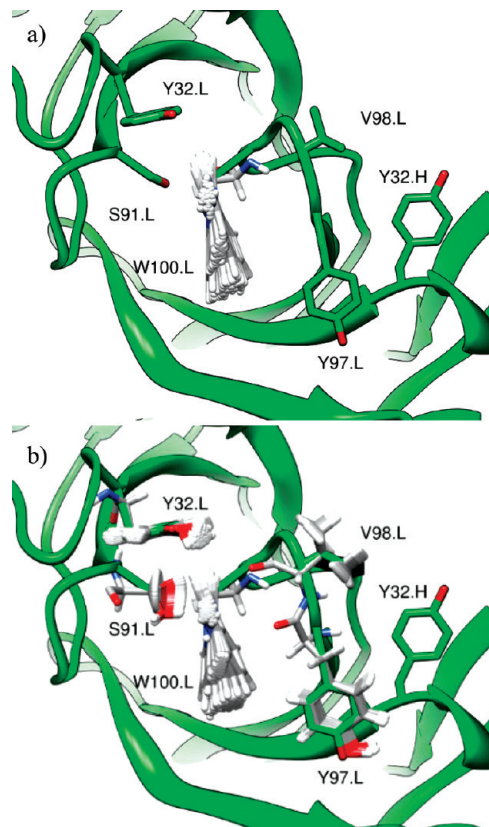


Figure 7. (a) Side chains of 1dba used for trajectories.. All remaining side chains and backbone configurations are held at the native state. (a) For single side chain case in Table 2, only TRP 100.L (in green) is sampled. (b) Multiple side chains sampled as reported in Table 3.

is observed, which tapers off at $N_I = 39$, since the cost of a trial move is 80 times that of a standard simulation. The total number of accepted outer steps is a good measure of the effectiveness of the sampling, and is shown in Figure 8c. The slight improvement in error at $N_I = 39$ occurs even though there are fewer newly accepted configurations because the trial moves are less correlated, which will also serve to improve the statistics.

For $N_G = 16$ (the energy landscape most densely populated with energy minima), the standard sampling approach is more effective. Put simply, the basins are sufficiently densely packed that a posh scheme is no longer needed. Note also that the number of newly accepted outer steps is much higher for the standard setting ($N_I = 0$). It is clear from these initial results that, as expected, the posh MC scheme will be most effective for energy landscapes with relatively few low energy basins separated by large energy barriers, as is frequently the case in macromolecular applications.

These results also suggest that the initial perturbation should be large relative to the annealing step, and large enough to cross barriers, while the size of the annealing step is chosen to give good acceptance in sampling the local basin. It is usually straightforward to estimate the size of the annealing step. Often, the length scales emerge quite naturally from knowledge of the system. For example, a typical range for a single dihedral perturbation in a proteins or small molecules is typically less than $2\pi/3$, which is roughly the width of a single χ well.

Table 2. Energy Statistics of Single Side Chain Trajectory^a

	$N_I = 1$			$N_I = 5$	
	std	50% posh	100% posh	50%	100% posh
$N_O(\text{total}) / 10^4$	100	50	35	20	15
$\beta(\langle E \rangle - \langle E \rangle_0)$	0.07(0.06)	0.06(0.06)	0.09(0.12)	0.07(0.11)	0.00(0.11)
	-0.02(0.06)	0.00(0.09)	0.01(0.14)	0.05(0.10)	0.00(0.08)
σ	1.05	1.05	1.10	1.10	1.00
	1.05	1.08	1.20	1.12	1.18
$(\langle E \rangle - \langle E \rangle_0) / \sigma_0$	-0.01	0.00	-0.01	-0.04	0.00
	-0.01	0.00	0.01	-0.04	0.00
	$N_I = 10$			$N_I = 20$	
		50%	100% posh	50%	100% posh
$N_O(\text{total}) / 10^4$		15	8	6	4
$\beta(\langle E \rangle - \langle E \rangle_0)$		-0.01(0.12)	-0.01(0.12)	0.15(0.23)	0.17(0.33)
		-0.10(0.08)	0.12(0.08)	0.43(0.17)	0.46(0.18)
σ		1.11	1.03	1.31	1.39
		1.13	1.34	1.69	1.78
$(\langle E \rangle - \langle E \rangle_0) / \sigma_0$		-0.01	-0.01	0.15	0.17
		0.08	0.10	0.37	0.39

^a TRP 100.L of protein 1dba is simulated at 350K, as described in text (see Figure 7a). Values in gray are from a simulation using the full reverse transition calculation, and values in white are from a simulation using the stationary transition approximation (eq 28). Statistics are computed using 10 trajectories started from an equilibrated structure as an initial condition. Values in parentheses are the variance of the mean energy over the 10 trajectories, σ is the time averaged standard deviation of the energy trajectory. $\langle E \rangle_0$ and σ_0 are chosen as a reference state from the standard simulations at $N_I = 5$ (highlighted in yellow and rose). The total number of outer steps N_O used are summed over each of the 10 trajectories of equal length.

Table 3. Energy Statistics of Multiple Side Chain Trajectory at 350 K^a

a)

	$N_I = 1$		$N_I = 5$	
	50% posh	100%	50% posh	100%
$N_O(\text{total}) / 10^4$	125	90	50	20
$\beta(\langle E \rangle - \langle E \rangle_0)$	0.58(0.69)	0.43(0.56)	0.14(0.24)	0.00(0.43)
σ	2.40	2.27	2.34	2.24
$(\langle E \rangle - \langle E \rangle_0) / \sigma_0$	0.26	0.19	0.06	0.00
	$N_I = 10$		$N_I = 20$	
	50% posh	100%	50% posh	100%
$N_O(\text{total}) / 10^4$	20	15	10	8
$\beta(\langle E \rangle - \langle E \rangle_0)$	0.01(0.29)	0.36(0.33)	0.85(0.66)	0.66(0.27)
σ	2.28	2.47	2.57	2.76
$(\langle E \rangle - \langle E \rangle_0) / \sigma_0$	0.00	0.16	0.38	0.30

b)

Average Values for Individual Standard Trajectories					
$(\langle E \rangle - \langle E \rangle_0) / \sigma_0$	1.03	0.13	0.67	0.40	0.62
	0.71	0.36	0.18	1.03	0.67

^a Side chains of 1dba as shown in Figure 7b. (a) Statistics are reported as in Table 2. All trajectories are generated with the stationary transition assumption. (b) Average values of individual standard trajectories are reported for reference.

For the cases studied here, the acceptance ratios for the standard protocols for $N_G = 4$ and 9 (those showing improvement) were less than 1%. The use of posh sampling can be thought of as a method for refining trial moves that would otherwise be inefficient. For a given landscape and perturbation protocol, there exists an optimal N_I , which decreases to 1 as the landscapes become less sparse.

Empirical Corrections: Stationary Transitions. A practical consequence of the reverse pathway estimation in complex systems is that the reverse transitions corresponding to a rejected trial have been observed to result in an error in the estimate the ratio of forward and reverse transitions. In particular, the reverse transition calculations

are most susceptible to error, as the construction of the reverse pathway will often pass through sterically hindered portions of configuration space, resulting in anomalously high energies. This region can often lead to numerical error in the estimate of the transition probabilities. In these cases, it has proven to be useful to introduce the following empirical correction to the reverse transition probability:

$$t_{N-k, N-k+1}^{(4,1)} = \delta(\mathbf{q}_{N-k}^S - \mathbf{q}_{N-k+1}^T) \text{acc}_{N-k, N-k+1}^{(4,1)} + \delta(\mathbf{q}_{N-k}^S - \mathbf{q}_{N-k}^S) t_{N-k, N-k+1}^{(2,3)} \quad (28)$$

where $\text{acc}_{N-k, N-k+1}^{(4,1)}$ is the Metropolis acceptance probability for the reverse transition. This relation simply uses the

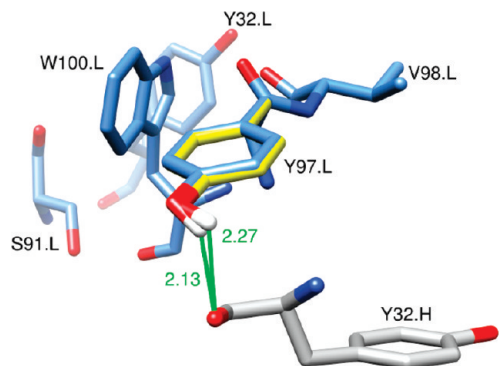


Figure 8. Anecdotal case of a large change in energy resulting from a small change in position. This change in position results in a $6.0k_B T$ change in energy during the course of a simulation.

forward transition probabilities in the reverse trial move when the forward trial is rejected, which prevents the overestimate of the ratio $T_{41}^{N_1}/T_{23}^{N_1}$, especially in longer inner loop settings. The motivation for the use of this correction comes from the notion of an ideal transition. If we consider the ideal forward transition move, it would consist of a series of purely downhill moves. Likewise, the reverse transition pathway would also consist of purely downhill moves. For these ideal transitions, the estimate of the ratio $T_{41}^{N_1}/T_{23}^{N_1}$ would be unity, if Metropolis transitions are used. The practical effect of estimating these ratios is to generate numbers that are slight corrections to this ratio. The introduction of the empirical correction helps to facilitate this assumption, and improved the statistics such that acceptable precision could be observed in the $N_1 = 20$ range. We have nominally assigned a tolerance of $\pm 0.1\sigma_0$ in the estimate of the average energy as an indicator for acceptable precision, where σ_0 is the standard deviation of the energy of a trajectory generated with a standard protocol. Table 2 compares simulations generated with and without the stationary transition assumption, and is discussed in the following section.

Molecular Applications

While a detailed study of model systems is worthwhile to understand the elements of the sampling, the ultimate goal is to apply this sampling to complex biomolecular systems. Here, we present two initial applications, specifically using posh Monte Carlo to generate ensembles for (a) side chains in an antibody binding pocket and (b) a phosphorylated tetrapeptide.

Side Chain Sampling: Precision Studies. Previous work has established the utility of side chain Monte Carlo sampling approaches to study binding pocket dynamics.⁴⁵ The published method perturbs a subset of protein side chain torsions while keeping the remaining backbone and side chain coordinates fixed. Here, we perform a similar simulation to study the effectiveness of the current algorithm in such a regime. The protein studied here is the Fab fragment of antibody protein 1dba,^{55,56} which was also considered in the previous study.⁴⁵

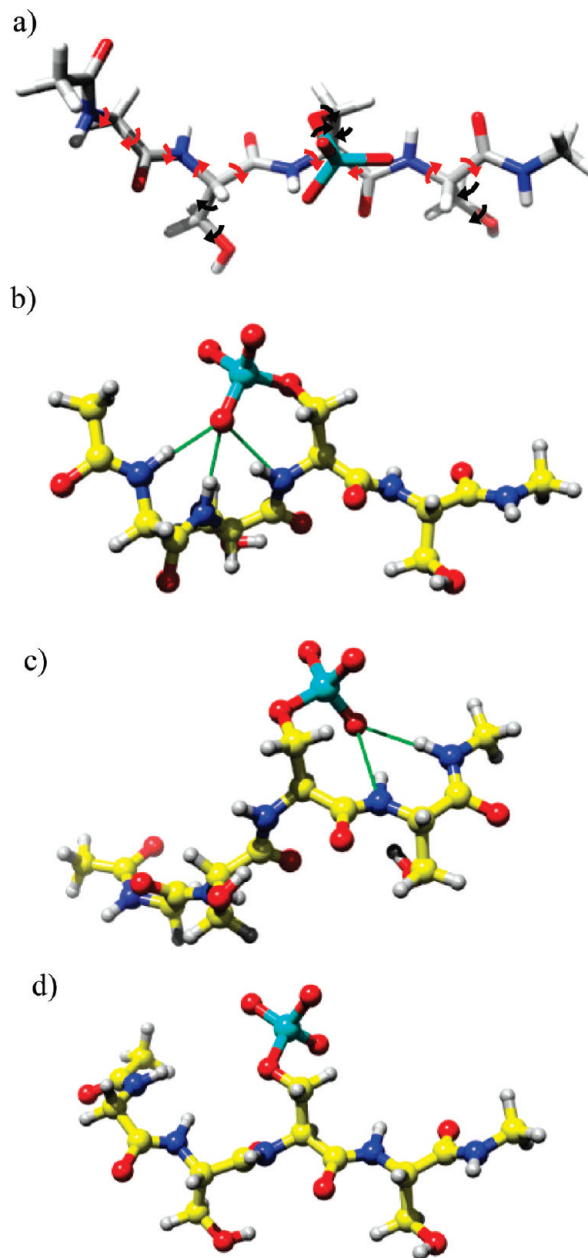


Figure 9. Phosphopeptide Ace-Gly-Ser-pSer-Ser-Nma studied. a) Dihedral angles sampled. Red arrows indicate all backbone dihedral angles allowed to fluctuate, while black arrows indicate the side chain dihedrals allowed to fluctuate. All other degrees of freedom are held fixed. (b, c, d) Typical hydrogen bonding configurations: (b) N-terminal hydrogen bonding network, (c) C-terminal hydrogen bonding network, and (d) solvated phosphate configurations.

To account for implicit solvation, the SGB model⁵⁷ is used, but for these cases, the Born Radii are held fixed at that which was calculated for the initial conformation. This constraint avoids introducing the additional bookkeeping of the Born Radii updates and is justifiable in this case because the solvent exposure of the side chains does not change dramatically as the side chains explore different conformations. This bookkeeping, however, will be reintroduced for the phosphopeptide case, where solvent exposure of various portions of the peptide change dramatically among sampled conformations, with the appropriate modified acceptance

Table 4. Energy Statistics of Phosphopeptide at 300 K^a

	$N_I = 1$			$N_I = 5$		
	std	50% posh	100%	std	50% posh	100%
$N_O/10^4$	1000	1000	1000	400	400	400
$\beta(\langle E \rangle - \langle E \rangle_0)$	0.51(0.70)	-0.15(0.60)	0.24(0.59)	0.00(0.41)	0.12(0.79)	0.26(0.67)
$\beta\sigma$	3.65	3.55	3.55	3.66	3.61	3.61
$(\langle E \rangle - \langle E \rangle_0)/\sigma_0$	0.14	-0.04	0.06	0.00	0.03	0.07
	$N_I = 10$			$N_I = 20$		
	std	50% posh	100%	std	50% posh	100%
$N_O/10^4$	200	200	200	100	100	100
$\beta(\langle E \rangle - \langle E \rangle_0)$	-0.64(0.38)	-0.58(0.40)	0.11(0.97)	-0.10(0.56)	-0.05(0.83)	1.62(0.64)
$\beta\sigma$	3.61	3.64	3.88	3.66	3.77	4.26
$(\langle E \rangle - \langle E \rangle_0)/\sigma_0$	-0.17	-0.16	0.03	-0.03	-0.01	0.44

^a Statistics are reported as described in Tables 2 and 3. All trajectories are generated with the stationary transition assumption.

Table 5. *J* Coupling Data^a

exptl	Wong et al. simulation	$N_I = 1$			$N_I = 5$			
		std	posh	50%	std	posh	50%	
Gly1	5.63(0.12)	4.12(0.03)	3.24(0.78)	3.01(0.22)	3.30(0.35)	3.25(0.19)	3.19(0.20)	3.39(0.25)
Ser2	6.65(0.12)	6.62(0.11)	7.63(0.21)	7.76(0.16)	7.68(0.20)	7.70(0.12)	7.57(0.21)	7.65(0.09)
PSer3	5.48(0.12)	5.73(0.10)	5.62(0.34)	5.62(0.45)	5.46(0.39)	5.64(0.23)	5.76(0.48)	5.60(0.25)
Ser4	6.93(0.12)	7.37(0.03)	7.46(0.06)	7.43(0.05)	7.42(0.05)	7.42(0.05)	7.45(0.07)	7.42(0.03)
		$N_I = 10$			$N_I = 20$			
		std	posh	50%	std	posh	50%	
Gly1		3.11(0.15)	3.50(0.44)	3.19(0.19)	3.32(0.19)	3.88(0.23)	3.18(0.12)	
Ser2		7.67(0.13)	7.53(0.14)	7.68(0.10)	7.71(0.10)	7.40(0.10)	7.67(0.12)	
pSer3		5.58(0.24)	5.54(0.40)	5.56(0.35)	5.81(0.30)	5.76(0.27)	5.82(0.34)	
Ser4		7.43(0.03)	7.42(0.07)	7.43(0.03)	7.44(0.05)	7.42(0.07)	7.44(0.07)	

^a Data set is the same as that used for Table 4.

probability. The nonpolar term of the solvation model is from the Levy and Gallicchio⁵⁸ model, and the force field is OPLS-AA 2005.^{59,60} Figure 9a shows the set of side chains sampled, which were chosen because they form direct interactions with the ligand in the holo structure and because they occupied somewhat different rotameric conformations in holo and apo structures of the antibody, suggesting that they may be flexible. The remainder of the protein is held at its native state. The initial perturbation consists of selecting a side chain randomly, assigning a uniform deviate of the torsion angles over the domain $[-\pi/2, \pi/2]$. The inner step trial moves follow the same procedure, but over the smaller domain $[-\pi/36, \pi/36]$. The reverse trajectory is generated using the concurrent pathway approach, where each reverse trial is generated using eqs 24 and 25.

To demonstrate that the sampling obeys detailed balance, a simple trajectory using only the Trp100 side chain is first generated (see Figure 7a). Table 2 provides energy statistics for a range of inner loop settings, $N_I = 1, 5, 10,$ and 20 . For each parameter setting, 10 trajectories were generated at 350 K. A standard trajectory was generated, where each trial move follows the initial perturbation, or (1→2) recipe. Both 100% posh trajectories and 50% posh trajectories were generated. For this series, we also studied the effect of introducing the stationary transition assumption (eq 28).

The metric used to evaluate precision is the difference in mean energies divided by the standard deviation from a

nominal baseline trajectory. The trajectory generated using 100% posh ($N_I = 5$) was used as the baseline. A cutoff of ~ 0.1 to 0.15 was assigned to determine if the trajectories were converging to the correct expectation value. Any value larger than this is deemed to be indicative of a systematic bias in the sampling. Table 2 shows that nearly all parameter settings presented here satisfy this requirement, and that the system clearly obeys detailed balance for all settings up to $N_I = 10$. Clearly, the 50% posh sampling provides higher precision than 100% posh sampling, which is to be expected. The upper bound for maintaining reasonable precision appears to be roughly around $N_I = 20$. This upper bound appears to be consistent across all systems studied in this work, with the exception of the model systems, which do not show this limitation as prominently. We believe that the limitation appears in the biomolecular systems because the reverse trajectories often pass through extremely high energy regions of space. Statistics in this region are more likely to be error prone, whereas the model systems have better defined values throughout the domain of interest.

The use of eq 28 for the reverse transitions appears to reduce the error in the higher N_I regimes and not to affect the statistics in any noticeable way in the smaller regimes. Based on the observed success of this approximation, it is used for the remaining biomolecular systems studied. We believe that the origin of the error is the systematic overestimation of the reverse transition probability by

generating a trial move, as described in eq 25 (or eq 23 had the true reverse pathway been calculated). This error will become more apparent in the longer inner loop settings and thus will benefit more from the stationary transition approximation.

Side Chain Sampling: Efficiency. With the demonstration of precision established by the single side chain data of in Table 2, we study a slightly larger system to demonstrate the efficiency of the algorithm in a more complex system. The side chains sampled are shown in both yellow and green in Figure 7b, and all other settings are the same as in the single side chain case discussed above. Table 3 shows that most of the normalized differences in mean energies fall within an acceptable region, with the similar exception of $N_1 = 20$, as above.

The most striking result is in the anomalously high value (large error) of the standard trajectory (0.57), that is, not using posh. Notice also that the error in the measurement, 0.70, is also quite large relative to the other trajectories. Table 3b shows that of the 10 standard trajectories generated, only 2 are within 0.2 of the baseline. This does not indicate incorrect sampling of the posh trajectories. Rather, it indicates highly nonconverged sampling of the standard trajectories. The size of the initial trial move ($1 \rightarrow 2$) in both the posh and standard cases determines the accessibility of states, such that certain local minima would be inaccessible if a smaller trial move were used in either the standard or the posh case. The fair comparison to be made (as was done in the model system study of Figure 6), therefore, is a standard trajectory with a trial move comparable to the initial trial move of the posh trajectory.

There is, however, a similar, quasi-ergodic limitation that occurs if *only* large trial moves are used. The much finer detail of the landscape is not as well sampled, and yet the data set will appear to be converged. Consider an anecdotal case from a trajectory presented in Figure 8. This particular side chain displacement, consisting primarily of the movement of the polar hydrogen, results in a $6k_B T$ change in energy. While many of the hydrophobic fluctuations are generally of longer range in nature, the polar interactions tend to vary quite rapidly with position when in close proximity. This effect has been described in protein folding as the champagne landscape problem, and points to the necessity of combining large and small trial move sets. Here, we see that posh sampling is able to capture this effect, giving it a substantial advantage over the standard approach.

Phosphopeptide Sampling. As a second application, we applied posh to a phosphopeptide: Ace-Gly-Ser-pSer-Ser-Nma, which has been studied previously experimentally and computationally.^{61,62} This system is small but challenging for sampling. The phosphate group forms hydrogen bonds with different combinations of backbone amide NH groups, and transitions between these conformations are relatively slow, but accessible with standard (single replica) MD simulations.

For this study, only dihedral angles are allowed to fluctuate, excluding the peptide ω angles and capping methyl group torsions. Figure 9 shows all of the dihedral angles that are sampled. The SGB⁵⁷ implicit solvent model was also

used here, with an external dielectric of 80 and an internal dielectric of 2 (to crudely account for electronic polarizability). The phosphate partial charges are those used by Groban et al.,⁶³ based on a study conducted by Wong et al.⁶¹ An additional protocol, developed in previous work⁴¹ is added to optimize sampling, and is described below.

Implementation Details: Modular Perturbations. Perturbations are classified as either backbone moves or side chain moves. For the peptide, a backbone move consists of randomly selecting a single ϕ or Ψ dihedral angle, and assigning a uniform variate perturbation to that angle over a defined domain. For the initial trial perturbation, the domain is $[-\pi, \pi]$, while the inner step trial moves are over a smaller perturbation domain $[-\pi/10, \pi/10]$.

The side chain perturbation recipe here is slightly different than that for the antibody case discussed above. It follows the same prescription regardless of whether it is considered to be an initial trial or an inner step move. Here, each uniform variate is over the domain of either $[-\pi, \pi]$ or $[-\pi/20, \pi/20]$, determined randomly, with the selection probability of either domain given equal probability. This is the same protocol developed in previous work which studies only side chain fluctuations.⁴¹ The polar hydrogens are sampled over the domain $[-\pi, \pi]$ at every step for which that residue is selected.

At both the initial trial step generation and the inner step moves, either a backbone move or side chain move is selected with equal probability. This corresponds to the diagonal pathway, since all degrees of freedom are allowed to fluctuate in the ($1 \rightarrow 2$), as well as the ($2 \rightarrow 3$) transitions. To generate the ($4 \rightarrow 1$) transition, the concurrent reverse pathway is used (eqs 20 and 21). At each forward transition, a reverse trial move is generated using the same type of perturbation for the forward perturbation (side chain or backbone). Both forward and reverse trajectories use Metropolis transitions.

Multiple Time Step Monte Carlo Acceptance Probability. For this system, since we wish to compare to experimental NMR data, we introduce a Born radii correction term. To improve computational efficiency, some of the energy parameters (such as the Born radii) \mathbf{P}_L are held at the latent state of the original coordinate, giving the parameter set $\mathbf{P}_L(\mathbf{r}_I)$. Most notably, the long-range interactions and Born radii are held fixed during the inner loop sampling. At the end of each cycle of posh sampling, the resulting coordinate state is taken to be a trial move and subjected to the Multiple Time-Step Monte Carlo⁶⁴ acceptance criterion

$$\text{acc}_{I,F} = \min \left[1, \frac{p(\mathbf{r}_F | \mathbf{P}_L(\mathbf{r}_F))}{p(\mathbf{r}_F | \mathbf{P}_L(\mathbf{r}_I))} \right] \quad (29)$$

The probability $p(\mathbf{r}_F | \mathbf{P}_L(\mathbf{r}_I))$ is the Boltzmann factor of the energy evaluated at current coordinate state F with latent parameters from coordinate state I . For all cases studied here, a single posh cycle is followed by a parameter update. This criterion rigorously corrects for the error introduced by holding the latent parameters at a fixed state, such that the resulting distributions are equivalent to that of a sampling

protocol which updated these parameters at every state. This method has been developed and tested in a previous work.⁴¹ Here we refer to state *I* as the initial coordinate state and state *F* is the final coordinate. As in previous work,⁴¹ there is an optimal range of “inner” steps for this sampling protocol as well. Our interest in the present work, however, is only to maintain the rigorous description of the solvation parameters. As such, the number of inner steps before updating the latent parameters is N_I , where N_I is either the inner step setting during a push cycle, or it is the length of a standard trajectory between latent parameter updates.

To generate comparable trajectories without push sampling, a set of standard trajectories was also generated, which maintained the MTSMC sampling. For these trajectories, the same number of inner steps are used in between latent parameter updates. The trial moves for these are those which would be used for the (1 \rightarrow 2) moves in the equivalent push scheme.

Validation Results: Energy Statistics. To generate equilibrium data for validation and comparison to experiments, 10 trajectories were generated for each parameter setting, starting each from a different equilibrated initial condition. The parameter settings were inner steps settings of 1, 5, 10, and 20, with outer steps set to 1×10^6 , 4×10^5 , 2×10^5 , and 1×10^5 , respectively (to maintain the same number of energy evaluations among the simulations). Three sampling protocols were explored: (1) a standard procedure, which uses no push sampling but uses the MTSMC procedure; (2) a 100% push procedure, where a single push cycle using the inner step setting is used, followed by the MTSMC procedure; and (3) a 50% mixture of these two procedures, where a random selection of either of the two protocols is made at every outer step. It should be noted that the MTSMC protocol is used here in all cases.

Table 4 shows energy statistics for a range of inner step settings. As in the previous cases, quantitative agreement can be obtained for settings of up to $N_I = 10$, with systematic deviations appearing at $N_I = 20$, suggesting an upper bound for precision for this particular system. The reasons for this deviation are largely caused by the rugged energy landscape, as discussed above for the antibody test case.

It should be noted that the (4 \rightarrow 1) transitions for the peptide are likely to generate far more pathological energy configurations because of the sampling of the backbone degrees of freedom. This is apparent in a slight increase in deviation at $N_I = 20$ (0.44 versus ~ 0.2 to ~ 0.3 for the side chain cases).

Comparison to Experiment: Physical Observables. *J* couplings were computed using the Karplus equation⁶⁴

$$\langle^3J\rangle = \frac{1}{L} \sum_{i=1}^L (A \cos^2 \theta_i + B \cos \theta_i + C) \quad (30)$$

where $\langle J \rangle$ is the average NMR *J* coupling value, *L* is the length of the simulation, and θ_i is the dihedral angle of the *i*th snapshot of the H–N–C $_{\alpha}$ –H bond. We compare the computed values to experiments^{61,62} and previous simulations reported by Wong and Jacobson using molecular dynamics.⁶¹ Of all of the simulations, we expect the standard (MTSMC)

values to be most reliable because it contains data from all inner steps settings. We find good agreement between the standard simulation (with MTSMC) and experiment, most notably in the pSer coupling, which is the most sensitive to the phosphate interactions. We obtain poor agreement with experiment for the Gly coupling, but this anomaly is in part because *J* couplings are not well-defined with regard to glycines.

The agreement between our results and the previous results obtained using the molecular dynamics simulations⁶² is good, considering that a different forcefield (AMBER) and implicit solvation model were used. We obtain good agreement of the *J* couplings with the standard (MTSMC) simulation as a control, and notice also that the observables appear to be reasonable across a broader range of inner step settings than that observed with the control. This is because the comparison of the standard protocol to the push protocol is, in general, a more sensitive measure of precision than the experimental comparisons. The experimental measures, however, help to evaluate the complete design, which includes the forcefield, the implicit solvation model, and the geometric assumptions used, and establish a precedent for the overall accuracy of the approach. It should be noted here that satisfactory sampling is easily obtained with a standard Monte Carlo protocol, such that we do not see a marked improvement in sampling efficiency here with the introduction of the push sampling approach. This is not entirely surprising, considering that previous workers were able to sample this molecule with conventional dynamical methods. While the phosphopeptide makes an excellent system for evaluating precision, the side chain trajectories are more indicative of the capacity to improve the quality of sampling in more complex systems.

Conclusions and Future Directions

We have presented a sampling protocol designed to allow efficient sampling of sparsely distributed basins, such as those that are encountered in complex biomolecular energy landscapes. The protocol obeys detailed balance to the extent that the transition probabilities are correctly estimated. Three variants were presented in terms of pathways: adiabatic, hybrid, and diagonal. All of these approaches could conceivably be useful, depending of the system of interest. Each pathway provides adequate performance, but the diagonal and hybrid approaches are more robust for longer inner loop protocols. Two reverse pathway constructions were presented, and appear to be similar in terms of precision, with a slight preference given to the concurrent reverse pathway due to the ease of implementation. A study of efficiency in simple model systems suggests that the algorithm will be most efficient when considering very sparsely distributed local minima.

Sampling of densely packed side chains appears to benefit greatly from the introduction of push sampling because of the nature of the energy landscape, characterized by sparsely distributed minima with steep basins. An implementation in a more realistic small peptide system shows that comparison to experimental results gives reasonable agreement. The performance of the approach in complex systems is largely

limited by the nature of the reverse pathway construction, which is much more numerically challenging in a system with steric occlusions. It is also a costly calculation, which will place an upper bound on efficiency. Future work in this area should include methods for estimating the reverse transitions without the need to compute a reverse trajectory.

Acknowledgment. We are grateful to Chris McClendon and Pradipta Bhandyopadhyay for many helpful conversations and testing. This work was supported by NSF Grant MCB-0346399, and in part by NIH Grant R01-GM081710. M.P.J. is a consultant to Schrödinger LLC.

Appendix A: Model System

Initial evaluation of the posh sampling is performed using a simple 2 dimensional model. Consider a 2 dimensional Gaussian distribution

$$g_i(x, y) = w_i \exp(-u_i(x, y)) \quad (31)$$

where x and y are scalar variables, and the index i refers to the i th parameter set defining the distribution. Models of this sort have been used to analyze biomolecular systems because of the simple evaluation of the configuration integral and other observables.^{65,66} The potential is simply a harmonic potential^{68,69}

$$u_i(x, y) = \gamma_i [x - (x_0)_i - (y_0)_i \mathbf{R}_{\theta_i}^T \begin{bmatrix} (\sigma_1)_i^{-2} & 0 \\ 0 & (\sigma_2)_i^{-2} \end{bmatrix} \mathbf{R}_{\theta_i} \begin{bmatrix} x - (x_0)_i \\ y - (y_0)_i \end{bmatrix}] \quad (32)$$

where γ is a force constant, x_0 and y_0 are the coordinates of the minimum, and the σ_1 and σ_2 are the standard deviations along the principal axes of the distribution. \mathbf{R}_{θ} is a matrix which rotates the principal axes of the distribution by an angle θ

$$\mathbf{R}_{\theta_i} = \begin{bmatrix} \cos \theta_i & \sin \theta_i \\ -\sin \theta_i & \cos \theta_i \end{bmatrix} \quad (33)$$

The entire set of parameters defining a single Gaussian term, then, are given as

$$\mathbf{p} = (x_0 \ y_0 \ \sigma_1 \ \sigma_2 \ \theta \ w \ \gamma) \quad (34)$$

One advantage to using potentials of this type is that the free energy of each basin can be computed analytically. Also, potentials of mean force can be analytically computed. Potentials can be superimposed to generate nonlinear behavior:

$$\begin{aligned} G(x, y | \{\mathbf{p}\}_{N_F}) &= \sum_{i=1}^{N_F} g_i(x, y) \\ U(x, y | \{\mathbf{p}\}_{N_F}) &= -\ln(G(x, y | \{\mathbf{p}\}_{N_F})) \end{aligned} \quad (35)$$

For all cases presented, the temperature is unity.

Appendix B: Model System Studies

Model System Properties. Since the algorithm allows for the possibility of sampling two subspaces hierarchically, a

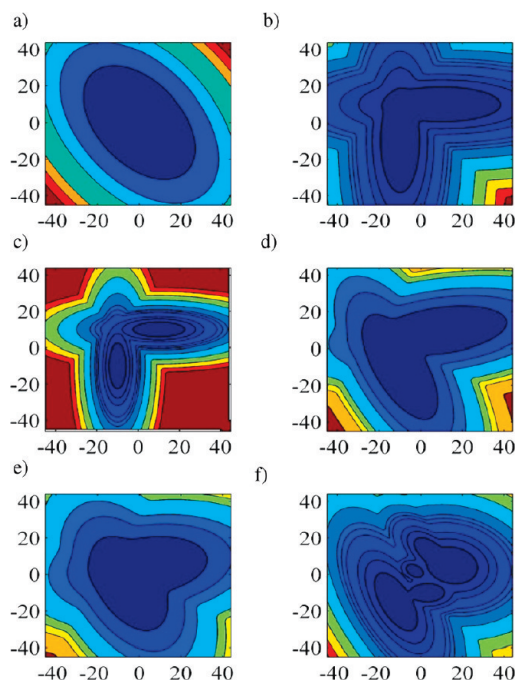


Figure 10. Landscapes studied. 2D contours of landscapes as described in text with parameters given in Table 4. For all plots, x is along horizontal axis. (a) Single basin. (b) Orthogonal disjoint basins. (c) Orthogonal accessible basins. (d) Acute basins. (e) Acute basins. (f) Multiple basins.

two-dimensional surface becomes a convenient model system for studying the precision of sampling (Figure 10). These model systems are superpositions of 2D Gaussian distributions, whose partition functions are known analytically. Table 6 lists the parameters used. These landscapes were designed to have features that are thought to influence the sampling. One of the features is the symmetry/asymmetry of the surface, which can play a role in the precision due to the symmetry in the reverse path construction. Another feature of interest is the distribution of local minima, either as quasi-ergodic or accessible basins. Landscapes with anharmonic saddle points and complex features were also included to mimic rough landscapes in a way that is intended to make errors in sampling more easily detected.

Choice of Perturbation Scheme. In applying the posh sampling to a system of interest, many considerations drive the choice of subspace partitions, but the choice will be limited to the types listed in Figure 2. It is worthwhile to notice the design considerations involved in each of the perturbation types. To explore this issue, we use a variety of simple two-dimensional energy landscapes, shown in Figure 5.

For Figure 11, the MSE of the adiabatic pathway with $N_1 = 1$ is defined as χ_0^2 and calculated for each of the pathways. The MSE, $\chi^2(N_1)$, for each value of N_1 is computed and the normalized value $\chi^2(N_1)/\chi_0^2$ is a measure of the relative error. Table 1 details the simulation conditions of the landscapes studied, as shown in Figure 10. The diagonal sampling scheme emerges as the most accurate of approaches across all landscapes shown here. This is expected because the sampling is unconstrained in both the perturbation and annealing steps. The hybrid approach also performs ex-

Table 6. Parameters of Landscapes as Described in Appendix A^a

landscape	x_0	y_0	σ_1	σ_2	w	θ/π	γ
A	0	0	3	5	1	0.2	0.2
B	-10	-10	4	12	1	0	9
	10	10	12	4	1	0	9
C	-10	-10	4	12	1	0	1.8
	10	10	12	4	1	0	1.8
D	-10	10	4	10	1	0.15	1
	10	10	4	10	1	0.55	1
E	-5	5	5	9	1	0.15	1
	5	5	5	9	1	0.55	1
F	-12	-15	6	14	3	0.15	6
	5	12	5	10	4	0.25	9
	3	-10	5	3	2	0	3
	15	5	10	8	3	-0.1	6
	-3	2	2	3	1	0.3	3
G	XX	XX	3	3	1	0.0	9

^a Letters correspond to those plotted in Figure 6. Landscape G describes a single basin of multiple basin landscapes which is plotted in Figure 5a.

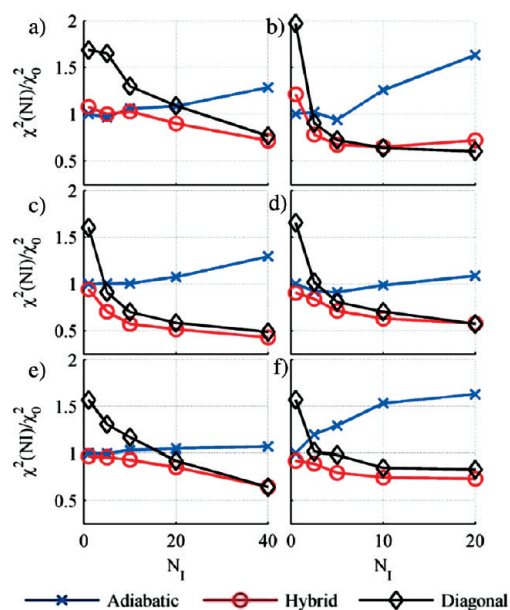


Figure 11. Comparison of errors for different pathway types. Errors are normalized to the Adiabatic case with $N_1 = 1$ for each landscape. Simulations were run with $N_0 = 40k$ steps. See Tables 1 and 4 for additional settings.

tremely well, suggesting that allowing all degrees of freedom to fluctuate in the annealing steps is a key to improving precision. The hybrid pathway often outperforms the diagonal pathway for smaller inner step settings, but this may be the result of the the initial perturbation of the hybrid case being effectively much smaller, since it only spans 1 dimension. At larger values of N_1 , the hybrid and diagonal pathways perform nearly identically. While the adiabatic pathway performs very well for smaller values of N_1 , it exhibits systematic bias as the number of inner steps is increased, due to the constrained sampling of the annealed coordinate. For shorter trajectories ($N_1 < 10$ for the cases here), the difference in error is negligible for all pathways. Any of these approaches, for relatively short trajectories, could be applied to more complicated systems with confidence, but in cases where longer inner loop trajectories are needed, the hybrid and diagonal pathways are more robust.

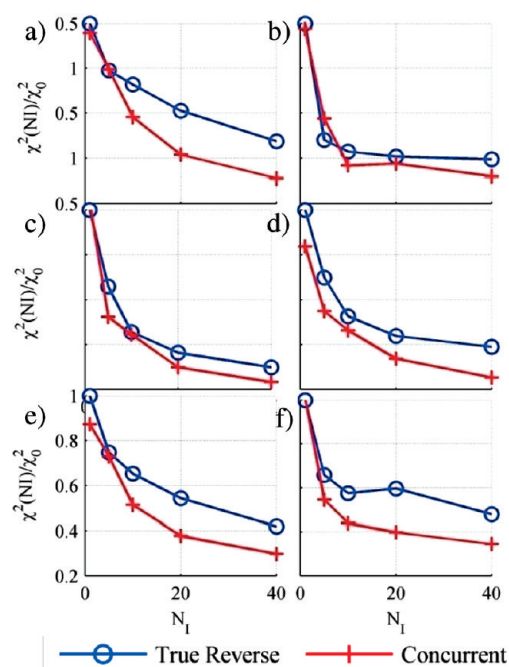


Figure 12. Accuracy of true and concurrent reverse paths are nearly equivalent. Each letter corresponds to the landscapes shown in Figure 5. MSE errors are normalized to the True Reverse pathway with $N_1 = 1$ for each landscape. Simulations were run with $N_0 = 40k$ steps. See Tables 1 and 4 for additional settings.

Choice of Reverse Pathway Construction. The estimate of the transition probability using the forward trajectory information is a common and relatively well understood method for estimating the forward transition probability.⁴⁷ Since the reverse pathway construction is the novel feature of this approach, it is worthwhile to compare the reverse pathway construction methods. Figure 12 shows the MSE (eq 22) of the landscapes shown in Figure 10. For Figure 12, the MSE of the true reverse pathway with $N_1 = 1$ is defined as χ_T^2 and a ratio $\chi^2(N_1)/\chi_T^2$ is calculated for each of the pathways. For all cases, the relative MSE steadily decreases with the number of inner steps. The concurrent reverse pathway appears to generate slightly smaller MSEs. The

performance across this range of landscapes suggests that either choice would be sufficient. This is somewhat surprising, since the shape of the reverse pathway is completely different for each case, and it is tempting to think that the shape of the landscape could have a more profound effect on the error. This does not appear to be the case, however, in these test cases. Since the concurrent reverse pathway is both reliable and easier to implement, it is likely to be used more widely, and is used exclusively for the biomolecular applications presented here.

References

- (1) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A.; Teller, E. *J. Chem. Phys.* **1953**, *21*, 1087.
- (2) Panagiotopoulos, A. Z.; Quirke, N.; Stapleton, M.; Tildesley, D. J. *Mol. Phys.* **1987**, *63*, 527.
- (3) Panagiotopoulos, A. Z. *Mol. Phys.* **2002**, *100*, 237.
- (4) De Pablo, J. J.; Prausnitz, J. M. *Fluid Phase Equilib.* **1989**, *53*, 177.
- (5) Rosenbluth, M. N.; Rosenbluth, A. W. *J. Chem. Phys.* **1955**, *23*, 356.
- (6) Frenkel, D.; Siepman, J. I. *Mol. Phys.* **1992**, *75*, 59.
- (7) Madras, N.; Sokal, A. *J. Stat. Phys.* **1988**, *50*, 109.
- (8) Hastings, W. K. *Biometrika* **1970**, *57*, 97.
- (9) Meirovitch, H. *J. Phys. A* **1982**, *15*, L735.
- (10) Meirovitch, H. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 9235.
- (11) Chen, B.; Siepman, J. I. *Theor. Chem. Acc.* **1999**, *103*, 87.
- (12) Chen, B.; Xing, J.; Siepman, J. I. *J. Phys. Chem. B* **2000**, *104*, 2391.
- (13) Opps, S.; Schofield, J. *Phys. Rev. E* **2001**, *63*, 56701.
- (14) Brown, S.; Head-Gordon, T. *J. Comput. Chem.* **2003**, *24*, 68.
- (15) Frantz, D. D.; Freeman, D. L.; Doll, J. D. *J. Chem. Phys.* **1990**, *93*, 2769.
- (16) Xu, H.; Berne, B. J. *J. Chem. Phys.* **1999**, *110*, 10299.
- (17) Zhou, R. H. *J. Chem. Phys.* **1997**, *107*, 9185.
- (18) Huber, G.; McCammon, J. *Phys. Rev. E* **1997**, *55*, 4822.
- (19) Swendsen, R. H.; Wang, J. S. *Physical Review Letters* **1986**, *57*, 2607.
- (20) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141.
- (21) Lyman, E.; Ytreberg, F.; Zuckerman, D. *Phys. Rev. Lett.* **2006**, *96*, 28105.
- (22) Roitberg, A.; Okur, A.; Simmerling, C. *J. Phys. Chem. B* **2007**, *111*, 2415.
- (23) Bandyopadhyay, P. *J. Chem. Phys.* **2008**, *128*, 134103.
- (24) Gordon, M.; Freitag, M.; Bandyopadhyay, P.; Jensen, J.; Kairys, V.; Stevens, W. *J. Phys. Chem. A* **2001**, *105*, 293.
- (25) Ulmschneider, J. P.; Ulmschneider, M. B.; Di Nola, A. *J. Phys. Chem. B* **2006**, *110*, 16733.
- (26) Li, Z. Q.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 6611.
- (27) David, L.; Luo, R.; Gilson, M. K. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 157.
- (28) Killian, B. J.; Yundenfreund Kravitz, J.; Gilson, M. K. *J. Chem. Phys.* **2007**, *127*, 024107.
- (29) Dill, K.; Chan, H. *Nat. Struct. Biol.* **1997**, *4*, 10.
- (30) Dill, K. *Biochemistry* **1990**, *29*, 7133.
- (31) Fiser, A.; Sali, A. *Methods Enzymol.* **2003**, *374*, 461.
- (32) Rohl, C. A.; Strauss, C. E.; Misura, K. M.; Baker, D. *Methods Enzymol.* **2004**, *383*, 66.
- (33) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. *J. Mol. Biol.* **1982**, *161*, 269.
- (34) Foreman, K. W.; Phillips, A. T.; Rosen, J. B.; Dill, K. A. *J. Comput. Chem.* **1999**, *20*, 1527.
- (35) Roux, B. *Implicit Solvent Models*; Marcel Dekker: New York City, 2001.
- (36) Rosso, L.; Mináry, P.; Zhu, Z.; Tuckerman, M. *J. Chem. Phys.* **2002**, *116*, 4389.
- (37) Rosso, L.; Tuckerman, M. *Mol. Simul.* **2002**, *28*, 91.
- (38) Darve, E.; Wilson, M.; Pohorille, A. *Mol. Simul.* **2002**, *28*, 113.
- (39) Coutsiar, E. A.; Seok, C. L.; Jacobson, M. P.; Dill, K. A. *J. Comput. Chem.* **2004**, *25*, 510.
- (40) Wedemeyer, W. J.; Scheraga, H. A. *J. Comput. Chem.* **1999**, *20*, 819.
- (41) Nilmeier, J.; Jacobson, M. *J. Chem. Theory Comput.* **2008**, *4*, 835.
- (42) Onuchic, J.; Luthey-Schulten, Z.; Wolynes, P. *Annu. Rev. Phys. Chem.* **1997**, *48*, 545.
- (43) Manousiouthakis, V. I.; Deem, M. W. *J. Chem. Phys.* **1999**, *110*, 2753.
- (44) Crooks, G. *J. Stat. Phys.* **1998**, *90*, 1481.
- (45) Jarzynski, C. *Phys. Rev. Lett.* **1997**, *78*, 2690.
- (46) Hetenyi, B.; Bernacki, K.; Berne, B. J. *J. Chem. Phys.* **2002**, *117*, 8203.
- (47) Jarzynski, C. Arxiv preprint cond-mat/0603185, 2006.
- (48) D Frenkel, B. S. *Understanding Molecular Simulation: From Algorithms to Applications*; Academic Press: Boston, 2002.
- (49) Dellago, C.; Bolhuis, P.; Csajka, F.; Chandler, D. *J. Chem. Phys.* **1998**, *108*, 1964.
- (50) Valleau, J.; Whittington, S. *J. Comput. Phys.* **1977**, *24*, 150.
- (51) Bolhuis, P.; Dellago, C.; Geissler, P.; Chandler, D. *J. Phys. Condens. Matt.* **2000**, *12*, 147.
- (52) Singhal, N.; Snow, C.; Pande, V. *J. Chem. Phys.* **2004**, *121*, 415.
- (53) Cao, J.; Berne, B. J. *J. Chem. Phys.* **1990**, *92*, 1980.
- (54) Zhou, R.; Berne, B. J. *J. Chem. Phys.* **1997**, *107*, 9185.
- (55) Arevalo, J. H.; Hassig, C. A.; Stura, E. A.; Sims, M. J.; Taussig, M. J.; Wilson, I. A. *J. Mol. Biol.* **1994**, *241*, 663.
- (56) Arevalo, J. H.; Stura, E. A.; Taussig, M. J.; Wilson, I. A. *J. Mol. Biol.* **1993**, *231*, 103.
- (57) Ghosh, A.; Rapp, C. S.; Friesner, R. A. *J. Phys. Chem. B.* **1998**, *102*, 10983.
- (58) Gallicchio, E.; Zhang, L. Y.; Levy, R. M. *J. Comput. Chem.* **2002**, *23*, 517.
- (59) Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. *J. Phys. Chem. B.* **2001**, *105*, 6474.
- (60) Rizzo, R. C.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1999**, *121*, 4827.

- (61) Wong, S. E.; Bernacki, K.; Jacobson, M. *J. Phys.Chem. B* **2005**, *109*, 5249.
- (62) Smart, J.; McCammon, J. A. *Biopolymers* **1999**, *49*, 225.
- (63) Groban, E. S.; Narayanan, A.; Jacobson, M. P. *PLoS Comput. Biol.* **2006**, *2*, e32.
- (64) Bernacki, K.; Hetenyi, B.; Berne, B. J. *J. Chem. Phys.* **2004**, *121*, 44.
- (65) Karplus, M. *J. Chem. Phys.* **2004**, *30*, 11.
- (66) Haliloglu, T.; Bahar, I.; Erman, B. *Phys. Rev. Lett.* **1997**, *79*, 3090.
- (67) Atilgan, A.; Durell, S.; Jernigan, R.; Demirel, M.; Keskin, O.; Bahar, I. *Biophys. J.* **2001**, *80*, 505.
- (68) Stern, H. A. *The Journal of Chemical Physics* **2007**, *126*, 164112.
- (69) Ballard, A. J. Jarzynski, C. *Proceedings of the National Academy of Sciences* **2009**.
CT8005166

Simple Expressions for Contact Values of Density Profiles in a Planar Double Layer

Douglas Henderson

*Department of Chemistry and Biochemistry, Brigham Young University,
Provo, Utah 84602*

Lutful Bari Bhuiyan*

*Laboratory for Theoretical Physics, Department of Physics, University of Puerto Rico,
San Juan 00931, Puerto Rico*

Received March 16, 2009

Abstract: Exact, or even accurate, results are very valuable for developing improved statistical mechanical theories of, for example, the double layer formed by the ions in an electrolyte near a charged electrode. In this paper we consider the primitive model of the double layer. For simplicity, we assume the electrolyte to be binary and symmetric. In our previous work, we proposed an exact local expression for the contact value of the total ion density, which is valid for all electrode charges, and local expressions for the individual ion density contact values for an electrode with a small charge. Previously, no local expression was known for the individual density contact values at high electrode charge. On the basis of our work and a recent paper by Lou and Lee [*J. Chem. Theory Comput.* 2009, 5, 1079–1083], we suggest local, empirical expressions for these contact values and test them by comparison with our extensive simulation results. The simulations are performed in the canonical ensemble with the ‘charged sheets’ method being utilized to treat long-range electrostatic interactions. The simulation results for the various contact quantities are seen to validate the empirical results for all electrode charges treated. The results are also likely to be (i) useful pointers for simulation of more complex, realistic models and (ii) relevant for experiments in that most electrochemistry occurs near the electrode where the classical Gouy–Chapman–Stern theory may not be adequate.

1. Introduction

If an electrode is inserted into an electrolyte, a charged layer is formed near the electrode. This layer is conventionally called the electric double layer even though modern work has shown that it may consist of multiple layers of differently charged ions. The study of this problem is both interesting and important because of its relevance for a plethora of industrial and biological systems. The electric double layer is commonly described using the primitive model (PM) in which the ions are treated as charge hard spheres and the solvent is treated as a dielectric continuum characterized by

the permittivity ϵ . If the ions are of equal size, this model is called the restricted primitive model (RPM). We treat here the RPM planar double layer and, for simplicity, restrict ourselves to double layers formed by binary symmetric $z:z$ valency electrolytes of charge ze , $-ze$, where z is the ion valence, e the magnitude of the electron charge, and d the common ionic diameter. The electrode is treated as a flat, smooth surface with a uniform charge density of σ and is assumed to have the same dielectric constant as the electrolyte so that polarization may be neglected.

The quantities of major importance in a theory of the electric double layer are the density profiles, $\rho_i(x) = \rho_i g_i(x)$, where i is the ion species, $\rho_i = \rho_i(\infty)$ is the bulk density of an ion of species i , and x is the perpendicular distance from

* To whom correspondence should be addressed. E-mail: beena@beena.cnet.clu.edu.

the electrode. The quantity $g_i(x)$ is the electrode–ion singlet distribution function. Both charge profiles and potential profiles may be obtained from the density profiles. The classical and traditional theory of the double layer is the seminal Gouy–Chapman–Stern (GCS)^{1–3} theory, which is based on the assumption that the ions are essentially free, devoid of mutual (interionic) correlations and that the ionic size may be neglected (except to define a distance of closest approach to the electrode). In the RPM, the distance of closest approach is $d/2$.

Exact, or at least accurate, expressions for the density or charge profiles are very valuable in gaining insight into the accuracy of a theory and developing new theories. A fundamental expression is the charge neutrality theorem in which the total integrated charge in the double layer is equal in magnitude but opposite in sign to the charge of the electrode. Although fundamental, it is not very helpful since any decent theory will satisfy this condition. Contact value theorems or exact sum rules are more helpful. Some years ago, Henderson and Blum⁴ and Henderson et al.⁵ obtained an exact contact value theorem for the total density profile $\rho(x) = \sum_i \rho_i g_i(x)$. For a binary symmetric electrolyte, we may define the sum and difference profiles by $g_{\text{sum}}(x) = [g_1(x) + g_2(x)]/2$ and $g_{\text{diff}}(x) = [g_1(x) - g_2(x)]/2$, respectively. The total density profile is obtained from $g_{\text{sum}}(x)$, while the charge profile is obtained from $g_{\text{diff}}(x)$. Without loss of generality, we assume the electrode to have a negative surface charge and define species 1 and 2 to be the counterions (+) and co-ions (–), respectively. With these definitions, the Henderson et al. theorem reads

$$g_{\text{sum}}(d/2) = a + \frac{b^2}{2} \quad (1)$$

Here $a = p/\rho k_B T$, with p , k_B , and T , being the pressure, Boltzmann constant, and temperature, respectively, $b = 4\pi\beta z e \sigma / \epsilon \kappa$ is a dimensionless variable for the electrode charge, and $\beta = 1/k_B T$. The parameter κ is the usual Debye–Hückel screening parameter and defined by $\kappa^2 = 4\pi\beta z^2 e^2 \rho / \epsilon$. As mentioned, eq 1 is exact and, thus, valid for all electrode charge. It is also local.

Until recently, eq 1 was the only contact value theorem for the double layer. What is needed is a contact theorem for $g_{\text{diff}}(x)$. Very recently, Holovko et al.^{6,7} obtained the exact contact value theorem for this quantity. For symmetric valency RPM electrolytes their expression can be written as

$$g_{\text{diff}}(d/2) = -\beta z e \int_{d/2}^{\infty} \frac{\partial \phi(t)}{\partial t} g_{\text{sum}}(t) dt \quad (2)$$

where $\phi(x)$ is the potential profile. Equation 2 is exact, valid for all electrode charge, but nonlocal and difficult to use.^{8,9} At about the same time, Henderson and Boda¹⁰ suggested an expression for $g_{\text{diff}}(d/2)$ at small electrode charge (small b). It is

$$g_{\text{diff}}(d/2) = ab + O(b^3) \quad (3)$$

This result was not derived; it was obtained from an examination of simulation results for a wide range of states. Thus, it cannot be regarded as exact but does seem accurate.¹¹ Additionally, it

is local. This result accounts for the low contact values of $g_{\text{diff}}(x)$ when the coupling constant $\beta z^2 e^2 / \epsilon d$ is large and a is small.

From eqs 1 and 3, the following local expressions for the $g_i(d/2)$ can be obtained

$$g_1(d/2) = g_+(d/2) = a + ab + \frac{b^2}{2} + O(b^3) \quad (4)$$

$$g_2(d/2) = g_-(d/2) = a - ab + \frac{b^2}{2} + O(b^3) \quad (5)$$

Bhuiyan et al.¹² pointed out that the product $g_+(x)g_-(x)$ is of interest because it is identically equal to 1 in the GCS theory. Hence, any departures of this product from unity give one insight into ways in which the GCS theory may be improved. From the above equations, we obtain the contact theorem

$$g_+(d/2)g_-(d/2) = a^2 + (a - a^2)b^2 + O(b^4) \quad (6)$$

which is accurate at small electrode charge.¹² An inherent difficulty with this relation is that at high electrolyte concentrations ($a > 1$), the second term on the right-hand side becomes negative, which in turn can lead to unphysical negative values for $g_+(x)g_-(x)$. Note that eqs 3–6 reduce to the corresponding GCS expressions when $a = 1$, the ideal gas limit. Clearly, accurate local expressions for $g_{\text{diff}}(d/2)$, $g_i(d/2)$, and $g_+(d/2)g_-(d/2)$ would be very useful. This is the subject of this note.

2. Theory and Simulations

2.1. New Expression. The biggest problem in generalizing eqs 5 and 6 is obtaining expressions that go to zero at large b without passing through negative values. Since the $g_i(x)$ are probabilities and normalized densities, they must be positive and, hence, $g_+(x)g_-(x)$ must be positive. Recently, Lou and Lee¹³ applied what they call a modified Poisson–Boltzmann theory to obtain an analytic expression for $g_+(x)g_-(x)$. In this theory the steric effect of the ions is taken into account by combining the classical mean field GCS theory with a lattice-gas type approximation for the ionic exclusion volume term. Since in the double-layer literature the term-modified Poisson–Boltzmann theory is used to denote a more elaborate but nonanalytic theory, we prefer to call the Lou–Lee theory a simplified, improved Poisson–Boltzmann theory. Note that the interionic correlations are still neglected in this theory. In any case, they obtain

$$g_+(d/2)g_-(d/2) = \exp(-\rho^* b^2) \quad (7)$$

where $\rho^* = \rho d^3$, $\rho = \sum_i \rho_i$. They claim that this result is accurate for large b .

Equation 7 provides a key to generalizing eq 6. We cannot just use the first two terms of the right-hand side of eq 6 as a prefactor to the exponential since, as indicated earlier, these may lead to negative values for $g_+(d/2)g_-(d/2)$. We suggest

$$g_+(d/2)g_-(d/2) = (a^2 + ab^2) \exp\left(-b^2 \frac{1 + \rho^* b^2/2}{1 + \alpha b^2}\right) \quad (8)$$

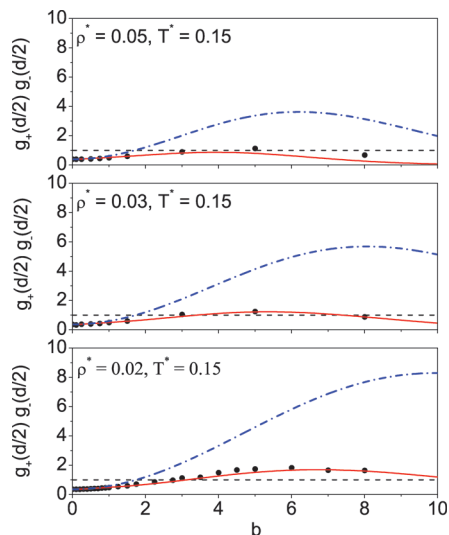


Figure 1. Product $g_+(d/2)g_-(d/2)$ as a function of b in a 1:1 RPM planar double layer at the reduced temperature $T^* = 0.15$ and reduced electrolyte density $\rho^* = 0.02$ (bottom), 0.03 (middle), and 0.05 (top). The symbols are the MC simulation data¹¹ while the solid and dash-dotted lines are the results from eq 8 with $\alpha = 1/2$ and 1, respectively. The dashed line represents the GCS result.

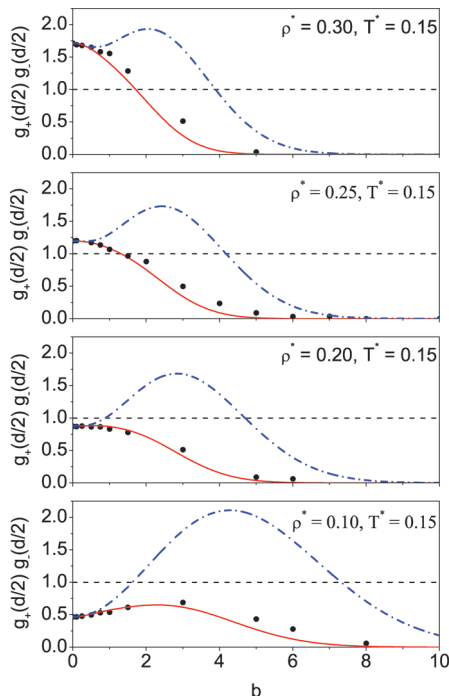


Figure 2. Same as Figure 1 but at reduced density: from bottom to top, $\rho^* = 0.1, 0.2, 0.25,$ and 0.3 .

where α is a parameter. Equation 8 reduces to eq 6 if b is small. When b is large, the argument of the exponential in eq 8 is $-b^2\rho^*/2\alpha$. We tried two values for α , 1 and $1/2$, with $\alpha = 1/2$ turning out to be the better choice. Note that at large b and $\alpha = 1/2$, eq 8 is consistent with the Lou–Lee expression (eq 7).

From eq 8 we obtain

$$g_{\text{diff}}(d/2) = \sqrt{[g_{\text{sum}}(d/2)]^2 - g_+(d/2)g_-(d/2)} \quad (9)$$

where $g_{\text{sum}}(d/2)$ is given by eq 1,

$$g_1(d/2) = g_+(d/2) = a + \frac{b^2}{2} + g_{\text{diff}}(d/2) \quad (10)$$

and

$$g_2(d/2) = g_-(d/2) = a + \frac{b^2}{2} - g_{\text{diff}}(d/2) \quad (11)$$

Because $g_+(d/2)g_-(d/2)$ and $g_+(d/2)$ are always positive, $g_-(d/2)$ is guaranteed to be positive, and eq 11 tends to zero as $b \rightarrow \infty$.

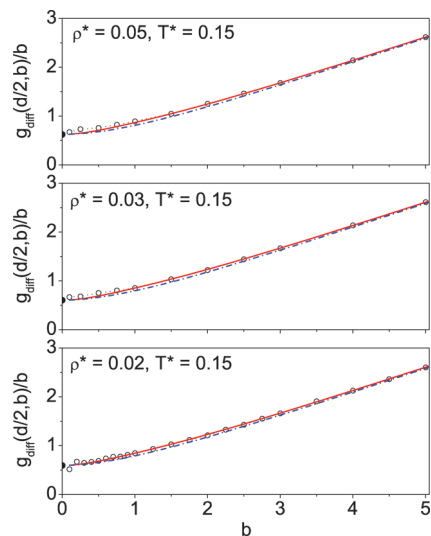


Figure 3. $g_{\text{diff}}(d/2, b)$ as a function of b in a 1:1 RPM planar double layer at the reduced temperature $T^* = 0.15$ and reduced electrolyte density $\rho^* = 0.02$ (bottom), 0.03 (middle), and 0.05 (top). The open circles are the MC simulation data,⁸ and the solid circle is the MC value of $a = g_{\text{sum}}(d/2, b = 0)$. The solid and dash-dotted lines are the results from eq 9, where the values $\alpha = 1/2$ and 1, respectively, have been used in $g_+(d/2)g_-(d/2)$.

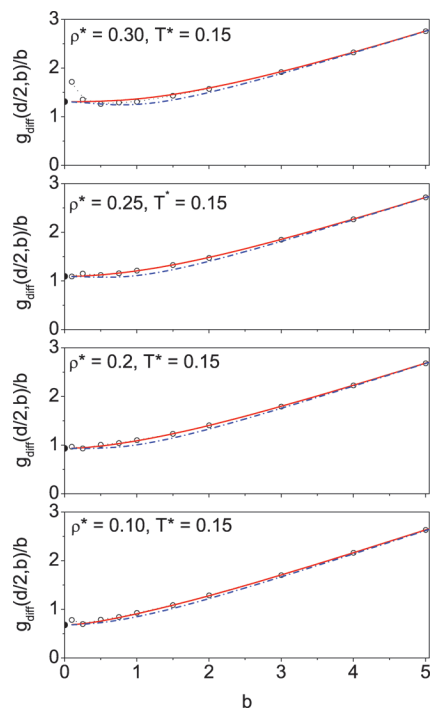


Figure 4. Same as Figure 3 but at reduced density: from bottom to top, $\rho^* = 0.1, 0.2, 0.25,$ and 0.3 .

2.2. Monte Carlo Simulations. The MC simulations were performed in the canonical (NVT) ensemble following the Metropolis algorithm. A rectangular parallelepiped of sides $l_x, l_y = l_z$ was chosen as the central MC box with one $y-z$ face being the planar charged wall at $x = 0$ and the other face at $x = l_x$ being uncharged. The traditional Ewald treatment of ionic interactions is cumbersome and computationally inconvenient for a planar system. In contrast, in such a geometry, the parallel *charged sheets* method pioneered by Torrie and Valleau¹⁴ to account for the long-range nature of the Coulomb interactions is attractive and convenient to use. The technique has subsequently been used successfully by other groups.^{15,16} Furthermore, in a recent work on electric double-layer capacitance Crozier et al.¹⁷ tested the charged sheets technique against a corrected 3D Ewald calculation using molecular dynamics and found the results from the two algorithms to be within the statistical error of each other for charged hard spheres. We employed the charged sheets method in conjunction with the usual minimum image plus periodic boundary conditions along the y and z axes. Note that Torrie and Valleau employed the grand canonical (μ V T) ensemble. The target bulk concentration was achieved through adjustments to the MC cell length l_x . The typical number of configurations sampled was on the order of 10^8 out of which the first 10^7 were used to equilibrate the system. The statistical uncertainty in reproducing the target bulk concentration was about $\pm 2\%$.

3. Results and Discussion

We present our empirical results for a range of physical states to test the predictions of eqs 8 and 9 against ‘exact’ simulation data. It is convenient to discuss results in terms of universal reduced parameters such as the reduced temperature $T^* = (\epsilon k_B T d) / (z^2 e^2)$ (reciprocal of the coupling parameter) and the reduced density $\rho^* = \rho d^3$. The specific values used are $T^* = 0.15$ and $\rho^* = 0.02, 0.03, 0.05, 0.1, 0.2, 0.25,$ and 0.3 . Although the reduced temperature $T^* = 0.15$ corresponds to a 1:1 solution at 75 K, for a 2:2 solution this represents room temperature, ~ 300 K. The values for the product $g_+(d/2)g_-(d/2)$ and $g_{\text{diff}}(d/2)$ from eqs 8 and 9, respectively, are compared with our previously published simulation results^{11,12} for the above range of states in Figures 1–4. In the figures results corresponding to both $\alpha = 1$ and $1/2$ are shown for comparative purposes.

The agreement of eqs 8 and 9 using $\alpha = 1/2$ is very good. Using $\alpha = 1$ consistently gives values for $g_+(d/2)g_-(d/2)$ that are too large relative to the simulations and show a maximum for all cases studied. The MC data show such maxima only for $a < 1$, and for $a > 1$ they decrease monotonically. Thus, for $\alpha = 1$, $g_+(d/2)g_-(d/2)$ is qualitative only at low electrolyte concentrations when $a < 1$. This feature notwithstanding (for $\alpha = 1$) the values for $g_{\text{diff}}(d/2)$ are reasonable and follow the simulation data quite closely. This suggests that we need to take a holistic approach rather than focus on individual quantities such as $g_{\text{diff}}(d/2)$. Equation 8 with $\alpha = 1/2$ is a very promising expression. We are fairly confident of eq 3; it is appealing

because it is consistent with the observed (in simulations) dewetting of the electrode at low charge and large coupling constant. However, lacking a more convincing justification, eq 8 can only be regarded as a useful, but empirical, local result. Nonetheless, there is no denying its accuracy. We note that obtaining sensible results is a nontrivial exercise. Because of the square root in eq 9, an imaginary result can be obtained easily.

A final remark or two on the empirical nature of the results is in order. Empirical results, per se are, in general, useful for assessing simulations and inspiring theoretical development. The simulation data are seen to validate the present results at all electrode charges treated, and the latter are thus likely to be useful pointers for simulations of more complex, realistic models. The various contact condition results presented here can be relevant for experiments also in view of the fact that most electrochemistry occurs at or near the electrode interface, the region where the classical GCS theory is not very accurate, being more a theory for integrated quantities.

Acknowledgment. L.B.B. wishes to acknowledge help from Dr. Jesus Piñero with regard to numerical analysis of some MC data.

References

- (1) Gouy, M. Constitution of the electric charge at the surface of an electrolyte. *J. Phys. (Paris)* **1910**, 9, 457.
- (2) Chapman, D. L. A contribution to the theory of electrocapillarity. *Philos. Mag.* **1913**, 25, 475.
- (3) Stern, O. The theory of the electrolytic double layer. *Z. Elektrochem.* **1924**, 30, 508.
- (4) Henderson, D.; Blum, L. Some exact results and the application of the mean spherical approximation to charged hard spheres near a charged hard wall. *J. Chem. Phys.* **1978**, 69, 5441.
- (5) Henderson, D.; Blum, L.; Lebowitz, J. L. An exact formulae for the contact value of the density profile of a system of charged hard spheres near a charged wall. *J. Electroanal. Chem.* **1979**, 102, 315.
- (6) Holovko, M.; Badiali, J.-P.; di Caprio, D. Contact conditions for the charge in the theory of the electrical double layer. *J. Chem. Phys.* **2005**, 123, 234705.
- (7) Holovko, M.; Badiali, J.-P.; di Caprio, D. New results from the contact theorem for the charge profile for symmetric electrolytes. *J. Chem. Phys.* **2007**, 127, 014106.
- (8) Henderson, D.; Bhuiyan, L. B. Contact value formulae for the density profiles of the electric double layer. *Collect. Czech. Chem. Commun.* **2008**, 73, 558.
- (9) Bhuiyan, L. B.; Henderson, D. Comment on ‘‘Contact conditions for the charge in the theory of the electrical double layer’’ [J. Chem. Phys. **123**, 234705 (2005)]. *J. Chem. Phys.* **2008**, 128, 117101.
- (10) Henderson, D.; Boda, D. On a conjecture of Fawcett. *J. Electroanal. Chem.* **2005**, 582, 16.
- (11) Henderson, D.; Bhuiyan, L. B. Evidence for a second contact value formula for the electric double layer. *Mol. Sim.* **2005**, 33, 953.

- (12) Bhuiyan, L. B.; Outhwaite, C. W.; Henderson, D. Some simulation and modified Poisson-Boltzmann theory results for the contact values of an electrolyte near a charged electrode. *J. Electroanal. Chem.* **2007**, *607*, 54.
- (13) Lou, P.; Lee, J. Y. Exact analytical result of contact value for the density in a modified Poisson-Boltzmann theory of an electrical double layer. *J. Chem. Theory Comput.* **2009**, *5*, 1079.
- (14) Torrie, G. M.; Valleau, J. P. Electrical double layers. I. Monte Carlo study of a uniformly charged surface. *J. Chem. Phys.* **1980**, *73*, 5807.
- (15) Boda, D.; Henderson, D.; Chan, K.-Y. Monte Carlo study of the capacitance of the double layer in a model molten salt. *J. Chem. Phys.* **1999**, *110*, 5346.
- (16) Lamperski, S.; Bhuiyan, L. B. Counterion layering at high surface charge in an electric double layer. Effect of local concentration approximation. *J. Electroanal. Chem.* **2003**, *540*, 79.
- (17) Crozier, P. S.; Rowley, R. L.; Henderson, D.; Boda, D. A corrected 3D Ewald calculation of the low effective temperature properties of the electrochemical interface. *Chem. Phys. Lett.* **2000**, *325*, 675.

CT900125Y

Frequency-Dependent Diffusion Constant of Quantum Fluids from Path Integral Monte Carlo and Tikhonov's Regularizing Functional

Piotr Kowalczyk,^{*,†} Piotr A. Gauden,[‡] Artur P. Terzyk,[‡] and Sylwester Furmaniak[‡]

Applied Physics, Royal Melbourne Institute of Technology University, GPO Box 2476 V, Victoria 3001, Australia and Department of Chemistry, Physicochemistry of Carbon Materials Research Group, Nicolaus Copernicus University, Gagarin St. 7, 87-100 Torun, Poland

Received May 1, 2009

Abstract: We present a novel implementation of the analytic continuation of the velocity autocorrelation function method that has been developed to study the transport properties of quantum liquids at finite temperatures. To invert the ill-posed linear Fredholm integral equation of the first kind, we combine Tikhonov's first-order regularizing functional with several methods used for automatic selection of the regularization parameter. Taking into account our results, we recommend two methods for automatic selection of the regularization parameter, namely: L-curve and quasi-optimality criterion. We found that the frequency-dependent diffusion power spectrum of normal liquid ^4He at $T = 4$ K and $\rho = 0.01873 \text{ \AA}^{-3}$ ($\rho = 31.1 \text{ mmol cm}^{-3}$) is characterized by a single asymmetric peak. The predicted self-diffusion coefficient of ^4He at this state point of $0.57\text{--}0.58 \text{ \AA}^2/\text{ps}$ is in excellent agreement with previous works. We demonstrate that, within proposed mathematical treatment of the quantum transport at finite temperatures, the entire real-time frequency-dependent diffusion power spectrum of liquid normal ^4He , can be successfully reconstructed from the limited number of Trotter slices and without the knowledge of covariance matrix. Moreover, the small values of regularization parameters (i.e., order of 10^{-7}) indicate that the information about quantum dynamics of normal liquid ^4He can be easily withdrawn from the high quality imaginary-time correlation function collected in the standard path integral Monte Carlo simulation.

Introduction

The understanding of dynamics of light particles (such as proton, electron, molecular hydrogen, helium, and others) at finite temperatures is essential for the designing of enzymatic reactions, solar cells, proton conducting membranes, quantum sieves, membranes for hydrogen purification, etc.^{1–7} Thus, the fundamental works for well-defined quantum systems at finite temperatures are necessary to validate the methods of quantum dynamics, where a satisfactory solution has not been found yet. Of all quantum liquids,

^4He is the most common and characteristic one.⁸ It is, therefore, an ideal system for investigating the microscopic origin of quantum effects and doing comparisons between experiments and theory.^{9–14}

Rabani et al.¹⁴ published the paper where they compared several techniques used for estimating the frequency-dependent diffusion constant in normal liquid helium above the λ transition. The techniques studied were the quantum version of mode-coupling theory (QMCT),^{15–18} the numerical analytic continuation method (NAC),^{14,19} the centroid molecular dynamics (CMD),^{20–24} and the Nakayama–Makri semiclassical approach (SC).^{25,26} Their study showed that all aforementioned methods yielded similar self-diffusion constants (i.e., zero time value of the velocity autocorrelation)

* Corresponding author. Telephone: +61 (03) 99252571. Fax: +61 (03) 99255290. E-mail: piotr.kowalczyk@rmit.edu.au.

[†] Royal Melbourne Institute of Technology University.

[‡] Nicolaus Copernicus University.

for normal liquid ${}^4\text{He}$ at $T = 4$ K and $\rho = 0.01873 \text{ \AA}^{-3}$ ($\rho = 31.1 \text{ mmol cm}^{-3}$). Moreover, all mentioned methods predicted asymmetric power spectrum of the quantum velocity autocorrelation function characterized by a single frequency peak. However, the power spectrum, over the entire frequency range, obtained with QMCT and NAC was quite different comparing to the results obtained using other methods. In other words, the diffusion of normal liquid ${}^4\text{He}$ at finite frequency (i.e., collective dynamics of quantum particles) depends on the method used for the study of dynamics. Hone et al.²⁷ investigated the same state point for ${}^4\text{He}$ using Feynman–Kleinert linearized path integral (FK-LPI) molecular dynamics,^{28,29} ring polymer molecular dynamics (RPMD),^{30–35} and CMD. They also observed significant differences among the results obtained for the quantum correction functions with these three methods over the short time and intermediate regions of the frequency spectrum. They concluded that it was not possible to assert which method was more accurate, owing to the lack of a benchmark result to compare with.

The analytic continuation of the velocity autocorrelation function is, in principle, an exact method that has been successfully used to study the transport properties of ‘realistic’ liquids.^{14–19,36} The imaginary-time correlation function that can be easily computed from the standard path integral Monte Carlo (PIMC) method yields a real frequency-dependent diffusion constant from an inversion of the integral equation. This problem is ill-posed (i.e., ill-conditioned) because the imaginary-time correlation function is defined over a finite range of imaginary time as well as it contains statistical noise.^{14–19,36–41} The Bayesian-based maximum entropy (ME) combined with the correlation matrix have been successfully used for inversion of the ill-posed integral equation. On the other hand, the limitations of ME have been reported.^{38–41} Numerous studies have shown that, in general, ME solutions tend to be too broad and smooth and often lack clarity defined peaks.^{38–41} In 1963, Tikhonov introduced a new method for approximate solution of ill-posed problems, known as a *regularizing algorithm*.⁴² Nowadays, Tikhonov’s regularizing functional, supported with several algorithms for automatic selection of the regularization parameter, is a very convenient method used for the approximate solution of board classes of ill-posed problems, including Fredholm and Volterra integral equations.^{43–46} It is worth nothing that stochastic regularization⁴⁷ (SR) as well as singular value decomposition³⁹ (SVD) have been used for study of the quantum many-body systems. However, these approaches have not met with much general success.

In the current paper, we implement the first-order Tikhonov’s regularizing functional for determination of the frequency-dependent diffusion constant at finite temperatures from the imaginary-time correlation function. Throughout numerical calculations, we show that the frequency-dependent diffusion constant can be successfully reconstructed from the limited number of Trotter slices and without the knowledge of covariance matrix. Moreover, for this particular integral equation, we recommended two methods for automatic selection of the regularization parameter, namely: L-curve and quasi-optimality criterion.⁴³ Finally, we com-

puted the frequency-dependent diffusion constant for liquid ${}^4\text{He}$ at 4 K and a density of $31.1 \text{ mmol cm}^{-3}$. As for numerical calculations, we found that the power spectrum of the normal liquid ${}^4\text{He}$ velocity autocorrelation function reconstructed from 64, 100, and 120 Trotter slices are almost identical. The predicted self-diffusion coefficient of helium at this state point of $0.57\text{--}0.58 \text{ \AA}^2/\text{ps}$ is in excellent agreement with previous works.^{14,27} The shape and the position of the frequency-dependent diffusion power spectrum agree well with the results of NAC and QMCT published by Rabani et al.¹⁴

Theory

A. Inversion of Frequency-Dependent Diffusion Constant from Tikhonov’s First-Order Regularizing Functional. Following Rabani et al.,^{14–19} the frequency-dependent diffusion constant, $D(\omega)$, can be computed from the following integral equation (from now on, we set $\hbar = 1$, $e = 1$, $M_e = 1$):

$$G_v(\tau) = \frac{1}{2\pi} \int_0^\infty [e^{-\omega\tau} + e^{-(\tau-\beta)\omega}] D(\omega) d\omega \quad (1)$$

where $G_v(\tau)$ denotes the imaginary-time correlation function, $\beta = (k_b T)^{-1}$ is the inverse temperature, $0 < \tau < \beta$ is the imaginary time, and P is the number of Trotter slices.

The imaginary-time correlation function can be computed in straightforward fashion using an appropriate PMIC simulation,^{14–19}

$$G_v(\tau_j) = \delta_{j1} \frac{1}{3m\varepsilon} - \frac{1}{N\varepsilon^2} \sum_{\alpha=1}^N \int d\mathbf{r}_1, \dots, d\mathbf{r}_p P(\mathbf{r}_1, \dots, \mathbf{r}_p) (\mathbf{r}_\alpha^j - \mathbf{r}_\alpha^{j-1}) (\mathbf{r}_\alpha^2 - \mathbf{r}_\alpha^1) \quad (2)$$

where $\varepsilon = \beta/P$, N is the total number of liquid particles, \mathbf{r}_j is a shorthand notation for the position vectors of all liquid particles assisted with bead j , \mathbf{r}_α^j is the position vector of liquid particle α of bead j , and $P(\mathbf{r}_1, \dots, \mathbf{r}_p)$ is the regular sampling function used in standard *cyclic* PIMC method (with $\mathbf{r}_0 = \mathbf{r}_p$).

Equation 1 is the *linear Fredholm integral equation of the first kind* with Laplace-type kernel, which is known to be ill-posed.^{42,43} We rewrite eq 1 in generic form:

$$Az = \int_0^b K(\tau, \omega) z(\omega) d\omega = G(\tau), \quad c \leq \tau \leq d \quad (3)$$

$K(\tau, \omega)$ is a real-valued function defined and continuous on the rectangle $\Pi = \{0 \leq \omega \leq b, c \leq \tau \leq d\}$.

We introduce the smoothing functional $M^\nu[z]$ (i.e., first-order Tikhonov’s regularizing functional):^{42,43}

$$M^\gamma[z] = \left\| Az - G \right\|^2 + \gamma \left\| z \right\|^2 = \int_c^d \left[\int_0^b K(\tau, \omega) z(\omega) d\omega - G(\tau) \right]^2 d\tau + \gamma \int_0^b \{ z^2(\omega) + [z'(\omega)]^2 \} d\omega \quad (4)$$

where $\gamma > 0$ denotes the *regularization parameter*.^{42,43}

We choose grids $\{\omega_j\}_{j=1}^n$ and $\{\tau_i\}_{i=1}^m$ on the intervals $[0, b]$ and $[c, d]$, respectively. For simplicity, it is assumed that the grids are uniform with steps h_ω and h_τ . Then, using trapezium quadrature we approximate eq 4 by finite-difference analog:

$$M^\gamma[z] = \sum_{i=1}^m \left[\sum_{j=1}^n a_{ij} z_j h_\omega - G_i \right]^2 h_\tau + \gamma \sum_{j=1}^n z_j^2 h_\omega + \gamma \sum_{j=2}^n \frac{(z_j - z_{j-1})^2}{h_\omega} \quad (5)$$

where

$$\begin{aligned} z_j &= z(\omega_j), & j &= 1, 2, \dots, n \\ G_i &= G(\tau_i), & i &= 1, 2, \dots, m \end{aligned} \quad (6)$$

$$\left. \begin{aligned} a_{ij} &= \frac{K(\tau_i, \omega_j)}{2}, & j &= 2, 3, \dots, n-1 \\ a_{ij} &= \frac{K(\tau_i, \omega_j)}{2}, & j &= 1, n \end{aligned} \right\} i = 1, 2, \dots, m \quad (7)$$

Note that for $\gamma > 0$, eqs 5–7 can be easily solved by any method of linear algebra because the resulting matrix is well-defined. However, the sole of the regularizing functional is a correct selection of the regularization parameter, γ , that controls the smoothness of the unknown function, $z(\omega)$. An increase in γ smoothes $z(\omega)$, but simultaneously the agreement between the $G_i(\tau)$ computed from PIMC and the theoretical one becomes worse. However, if γ is chosen to be small, then eq 5 would not be well-conditioned. Simply, the resulting matrix is still singular. The regularization parameter can be selected manually (i.e., through a series of trials by an interactive judgment of the solution) or automatically (i.e., applying some criteria). In the current work, we used three criteria for automatic selection of the regularization parameter, namely: the L-curve, the quasi-optimality criterion, and the generalized cross-validation method. However, we found that for considered regularizing functional, only L-curve and quasi-optimality criterion generate stable solutions.

Both L-curve and quasi-optimality criterion select the regularization parameter in inverse problems without taking into account the noise level. As mentioned by Hansen,⁴³ the most convenient graphical tool for analysis of discrete ill-posed problems is the so-called L-curve, which is a plot for all valid regularization parameters of the discrete smoothing norm $M^\gamma[z]$, e.g., the (semi)norm $\|z\|^2$ of the regularized solution versus the corresponding residual norm $\|Az - G\|^2$. The L-curve clearly displays the compromise between minimization of these two quantities, which is the heart of any regularization method. For discrete ill-posed problems, it turns out that the L-curve, when plotted in the log–log scale, very often has a characteristic L-shaped appearance (hence its name) with a distinct corner separating the vertical

and horizontal parts of the curve.⁴³ This corner corresponds to the optimal value of the regularization parameter. Following Tikhonov et al.,⁴² the optimal regularization parameter is determined from the minimization of the following norm,

$$\min_{\gamma \in \mathcal{R}^+} \left\| \gamma \frac{dz^\gamma}{d\gamma} \right\|^2$$

where z^γ denotes the solution of the discrete smoothing norm $M^\gamma[z]$ for $\gamma > 0$. This rule known as quasi-optimality criterion works remarkably well in practice.

B. Theoretical Study. To investigate the limitations and accuracy of the Tikhonov's regularizing algorithm with automatic selection of γ , we performed an inversion of eq 5 using theoretical imaginary-time correlation data (see eq 1). Since the reported $D(\omega)$ power spectrum for normal liquid ⁴He is characterized by a single asymmetric peak,^{14,27} we model this function by a linear combination of two Gaussian functions:

$$D(\omega) = \zeta f_1 + (1 - \zeta) f_2 \quad (8)$$

$$f_i = \frac{1}{\sqrt{2\pi}\delta_i} \exp\left[-\frac{(\omega - \omega_i)^2}{2\delta_i^2}\right] \quad (9)$$

where $0 \leq \zeta \leq 1$, ω_i and δ_i denote the mean and standard deviation of i -th Gaussian function, respectively. We computed $G_i(\tau)$ from eq 1 with parameters corresponding to normal ⁴He at 4 K. The Feynman path integral representation of quantum particle at finite temperature is an exact for infinite number of Trotter slices.^{48,49} Thus, we address the following question: *how the finite number of Trotter slices in isomorphic cyclic polymer impacts the reconstruction of the frequency-dependent diffusion power spectrum?* To answer this question we computed $G_i(\tau)$ for a different number of Trotter slices, i.e., 500, 100, and 50. Next we reconstructed those $D(\omega)$ power spectra from eq 5 using the singular-value decomposition method (SVD) and automatically selected γ from L-curve and quasi-optimality criterion.

C. Normal Liquid Helium. Dynamic properties of low-temperature quantum liquids have been always difficult to measure or to calculate. The dynamic properties of helium at 4 K have been extensively studied by different approximate methods of quantum dynamics.^{14,27} That is why we focus on the computation of the frequency-dependent diffusion constant for ⁴He at 4 K and a density of 31.1 mmol cm⁻³. In order to obtain the imaginary-time correlation function, that is the input required for the inversion of eq 5, we have performed a series of PIMC simulations in a NVT ensemble at the mentioned above state point. In our simulations, we used 256 helium atoms interacting via HFD-B2 He(4) potential due to Aziz et al.⁵⁰ We investigated a different number of Trotter slices, i.e., 1, 4, 8, 16, 32, 64, 100, and 128. For each studied system, we performed 2×10^6 Monte Carlo steps using the staging algorithm. The kinetic energy was computed from thermodynamic estimator.

Results and Discussion

In figure 1, we present the reconstruction of the theoretical $D(\omega)$ power spectra for normal liquid ⁴He at 4 K. Note that

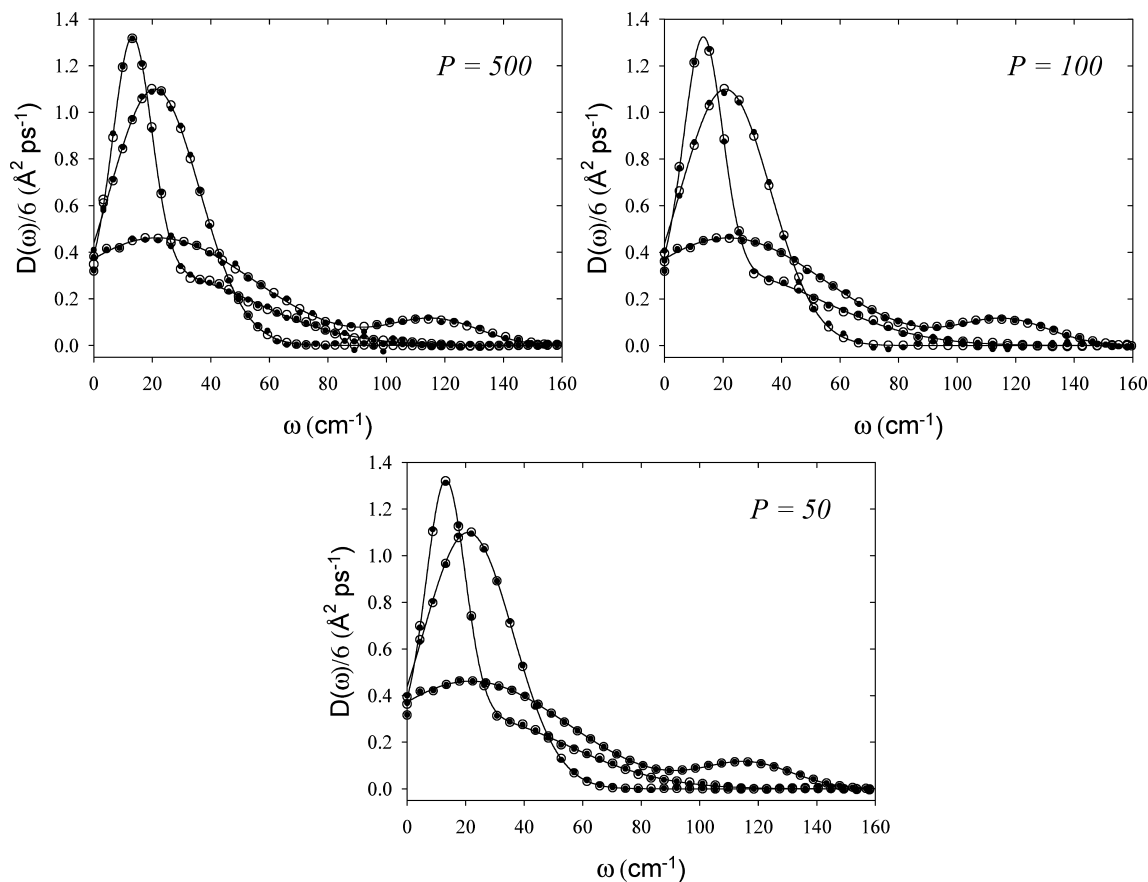


Figure 1. Reconstruction of the theoretical frequency-dependent diffusion power spectra (solid lines) from the first-order Tikhonov's regularizing functional using L-curve (open circles) and quasi-optimality criterion (close circles) for automatic selection of the regularization parameter. Theoretical curves are generated from eqs 1, 8, and 9 using parameters for normal liquid ${}^4\text{He}$ at 4 K and different numbers of Trotter slices, i.e., 500, 100, and 50.

we did not attach the random noise to numerical $G_v(\tau)$ generated from eqs 1, 8, and 9. This is because the high quality $G_v(\tau)$ can be computed from the PIMC simulation. Since $G_v(\tau)$ is always collected for the finite number of Trotter slices, we want to shed more light on the problem of incomplete imaginary-time correlation function. It is commonly known that the static properties of quantum fluids can be easily computed from the finite number of Trotter slices, once the kinetic energy estimator has been converged. However, as far as the dynamics of quantum fluids at finite temperatures is concerned, the same statement has not been reached yet. Throughout numerical experiments, we found that different positions as well as shapes of theoretical $D(\omega)$ power spectra are correctly reproduced using two methods for automatic selection of the regularization parameter, namely: L-curve and quasi-optimality criterion, as is presented in figure 1. Moreover, the first-order Tikhonov's regularizing functional implemented via eqs 5–7, is able to recover original entire $D(\omega)$ power spectra using different numbers of Trotter slices. An important conclusion of our theoretical calculations is that the $D(\omega)$ power spectrum can be successfully reconstructed from a limited number of Trotter slices and without the knowledge of covariance matrix. Let us now focus on the calculation of $D(\omega)$ power spectrum for normal liquid ${}^4\text{He}$ at 4 K and a density of $31.1 \text{ mmol cm}^{-3}$. First, we performed a series of PIMC simulations of ${}^4\text{He}$ at 4 K ($31.1 \text{ mmol cm}^{-3}$) to investigate

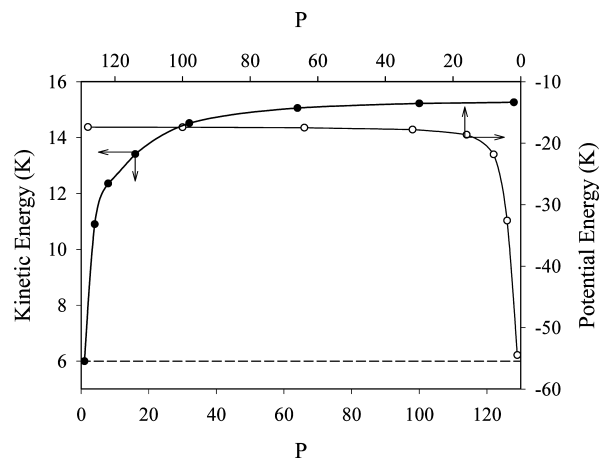


Figure 2. Kinetic and potential energy of normal liquid ${}^4\text{He}$ (4 K and density of $31.1 \text{ mmol cm}^{-3}$) as a function of the number of Trotter slices. Dashed line corresponds to the classical kinetic energy of ${}^4\text{He}$ at 6 K.

convergence of the kinetic and potential energy estimators (see figure 2). It is clear that convergence of kinetic and potential energy estimators have been achieved for the number of Trotter slices of 100. The kinetic energy of normal liquid ${}^4\text{He}$ at this state point of 15.3 K is close to 15.36 K reported by Hone et al.²⁷ In figure 3 we present $G_v(\tau)$ collected in PIMC simulation, where the helium atoms were quantized by 128 Trotter slices. Similarly to previous

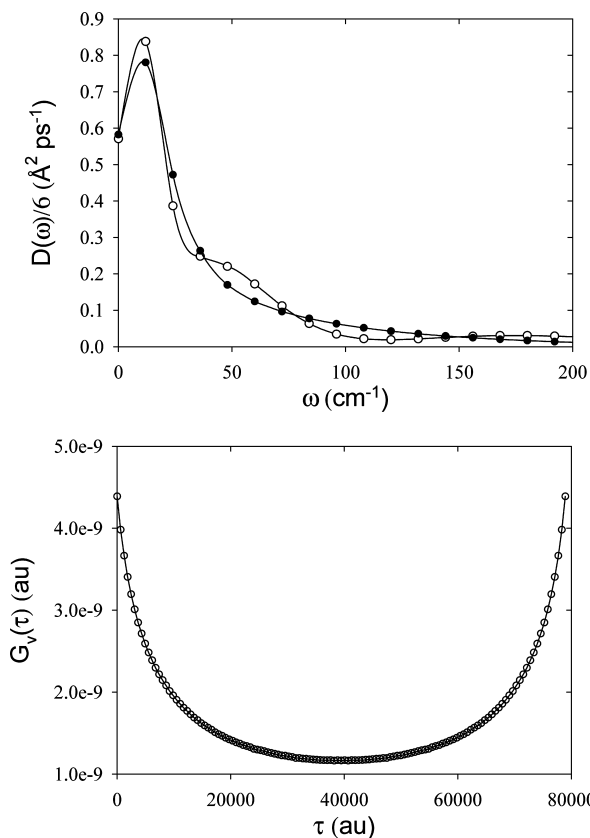


Figure 3. A plot of frequency-dependent diffusion constant for normal liquid ^4He at 4 K and density of $31.1 \text{ mmol cm}^{-3}$ computed from the first-order Tikhonov's regularizing functional using L-curve (open circles) and quasi-optimality criterion (close circles) for the automatic selection of regularization parameter. Bottom panel shows the imaginary-time correlation function computed from PIMC (open circles) and eq 1 (solid line).

reports,^{14,27} we found that the frequency-dependent diffusion power spectrum is characterized by a single asymmetric peak. Note, that both the L-curve and quasi-optimality criterion predict a similar value of the regularization parameter, as displayed in figure 4. The L-curve is characterized by a sharp well-defined corner with the optimum regularization parameter of 1.29×10^{-7} . The quasi-optimal criterion predicts a slightly higher value of the regularization parameter. That is why $D(\omega)$ computed from eq 5, using the second value of the regularization parameter, is smoother. This observation does not depend on the number of Trotter slices used in PIMC simulations (see figure 5). The small values of regularization parameters computed from the both criteria confirmed the high quality of $G_V(\tau)$ collected in our PIMC simulations. Strictly speaking, the final matrix given by eq 5 is singular for $\gamma = 0$; however, its singularity is easily removed by adding a small regularization penalty, i.e. $\gamma \approx 10^{-7} - 10^{-6}$. Predicted self-diffusion coefficient of liquid normal ^4He for this state point of $0.57 - 0.58 \text{ \AA}^2/\text{ps}$ is in excellent agreement with previous works.^{14,27} Let us now focus on the results presented in Figure 5. As for numerical experiments, we observe that $D(\omega)$ power spectrum can be successfully extracted from a limited number of Trotter slices (i.e., incomplete data of imaginary-time fluctuations). Indeed, for $P = 64, 100,$ and 128 , the computed $D(\omega)$ power spectra

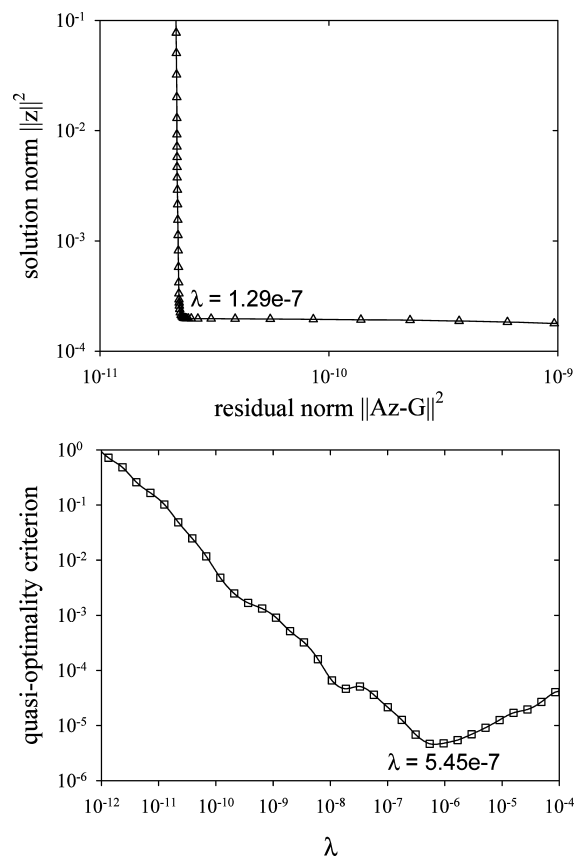


Figure 4. Regularization parameters computed from L-curve (top panel) and quasi-optimality criterion (bottom panel) for normal liquid ^4He (4 K and density of $31.1 \text{ mmol cm}^{-3}$) and the number of Trotter slices of 128.

are almost identical. Because the quasi-optimality criterion predicts higher values of regularization parameters in comparison to the L-curve, $D(\omega)$ power spectra inverted using these regularization parameters are smoother. However, both high and decay of the peaks are very similar for both regularization parameters. Surprisingly, even for $P = 32$ and regularization parameter selected from quasi-optimality criterion, the computed $D(\omega)$ power spectrum is very similar to those computed from higher number of Trotter slices.

Although, PIMC simulation generates incomplete data of imaginary-time correlation function (i.e., the number of Trotter slices quantizing quantum particle is in principle finite), our computational results indicate that real time $D(\omega)$ power spectrum of liquid normal ^4He at 4 K (in deep quantum region) can be reconstructed from this incomplete information. Furthermore, we show that within proposed mathematical treatment, the covariance matrix is not needed for the reconstruction of $D(\omega)$. Finally, small values of regularization parameters automatically selected from L-curve and quasi-optimality criterion indicate that the singularity of the final matrix given by eq 5 can be easily removed. Thus, it is not surprising that the unique $D(\omega)$ power spectrum of normal liquid helium is reconstructed from a limited number of Trotter slices and without the knowledge of covariance matrix. The extension of the proposed implementation of the analytic continuation of the velocity

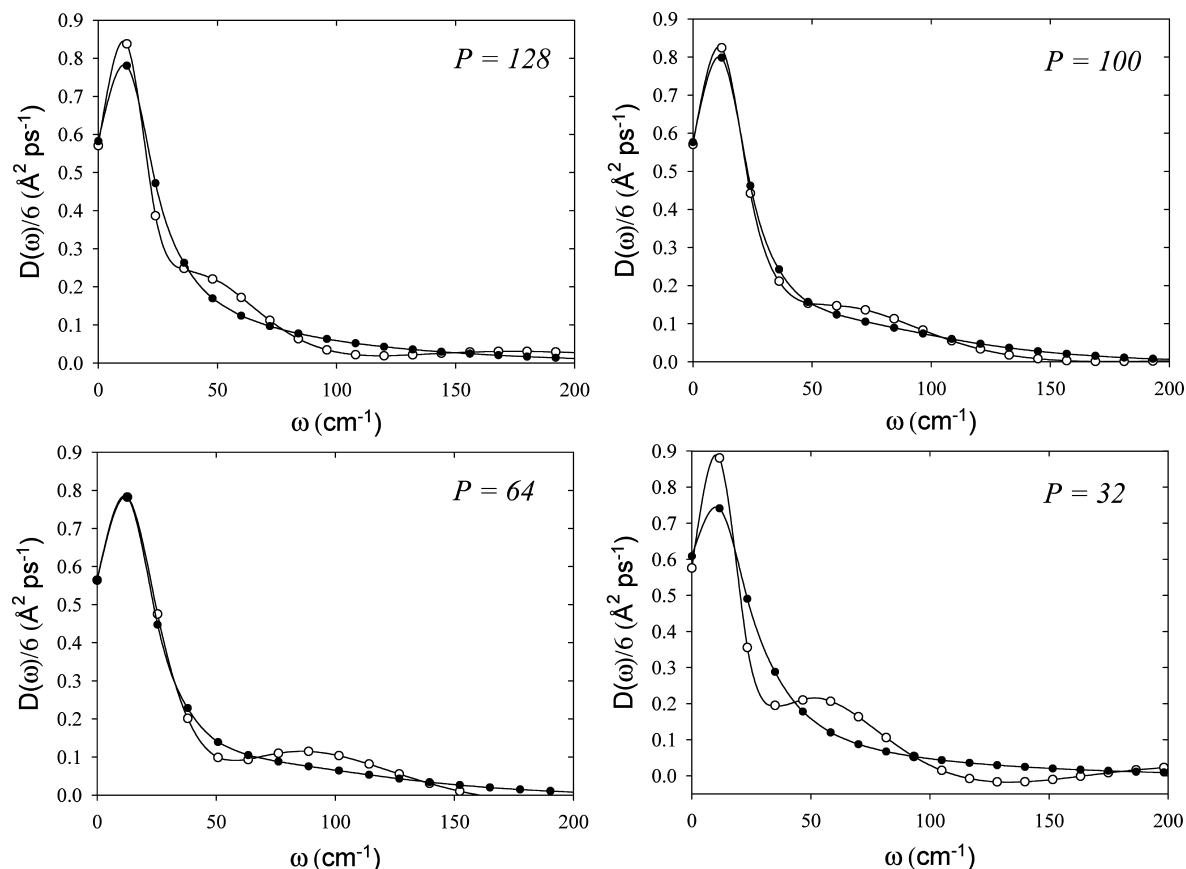


Figure 5. Plots of frequency-dependent diffusion power spectrum for normal liquid ${}^4\text{He}$ at 4 K and density of $31.1 \text{ mmol cm}^{-3}$ computed from the first-order Tikhonov's regularizing functional using L-curve (open circles) and quasi-optimality criterion (close circles) for automatic selection of the regularisation parameter. For PIMC simulations we used different number of Trotter slices, i.e., 128, 100, 64, and 32.

autocorrelation function to the other quantum systems will be the subject of our future works.

Conclusions

In the current work we present a novel implementation of the analytic continuation of the velocity autocorrelation function. This method allows computing a real-time frequency-dependent diffusion power spectrum of quantum fluid from the imaginary-time fluctuations collected in standard PIMC simulation. To solve the ill-posed Fredholm-type integral equation, we combine Tikhonov's first-order regularizing functional with several methods used for automatic selection of the regularization parameter. From our computational results we concluded that two methods used for automatic selection of regularization parameter, namely: the L-curve and quasi-optimality criterion predict reliable frequency-dependent power spectrum of liquid normal ${}^4\text{He}$. An important conclusion of this study is that, within proposed mathematical treatment of quantum dynamics at finite temperatures, a frequency-dependent diffusion constant can be successfully reconstructed from the limited number of Trotter slices and without the knowledge of covariance matrix. This remark seems to be important for study of transport of other light particles, such as an electron or proton near ambient temperatures.

Acknowledgment. P.K. would like to thank the discussions and hospitality of the Eran Rabani (School of Chemistry, Tel Aviv University, Israel). The authors acknowledge the use of the computer cluster VPAC (Victorian partnership for advanced computing). P.K. acknowledges the Royal Melbourne Institute of Technology for postdoctoral fellowship (Academic Level B, 2008-2010). P.G. acknowledges the use of the computer cluster at Poznan Supercomputing and Networking Centre as well as the Information and Communication Technology Centre of the Nicolaus Copernicus University (Torun, Poland). S. Furmaniak gratefully acknowledges the financial support from the Foundation for Polish Science.

References

- (1) Marx, D.; Tuckerman, M.; Hutter, J.; Parrinello, M. *Nature* **1999**, *397*, 601–604.
- (2) Kuki, A.; Wolynes, P. G. *Science* **1987**, *236*, 1647–1652.
- (3) Kamakoti, P.; Morreale, B. D.; Ciocco, M. V.; Howard, B. H.; Killmeyer, R. P.; Cugini, A.; Sholl, D. S. *Science* **2005**, *307*, 569–573.
- (4) Kowalczyk, P.; Gauden, P. A.; Terzyk, A. P.; Furmaniak, S. *J. Phys.: Condens. Matter* **2009**, *21*, 144210–144222.
- (5) O'Regan, B.; Graetzl, M. *Nature* **1991**, *353*, 737–740.
- (6) Schlapbach, L.; Züttel, A. *Nature* **2001**, *414*, 353–358.

- (7) Kowalczyk, P.; Gauden, P. A.; Terzyk, A. P. *J. Phys. Chem. B* **2008**, *112*, 8275–8284.
- (8) Van Sciver, S. W. *Helium Cryogenics*; Plenum Press: New York, 1986.
- (9) Nakayama, A.; Makri, N. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 4230.
- (10) Brualla, L.; Boronat, J.; Casulleras, J. *J. Low Temp. Phys.* **2002**, *126*, 1547.
- (11) Mazzanti, F.; Boronat, J.; Polls, A. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1996**, *53*, 5661.
- (12) Ceperley, D. M. *Mod. Phys.* **1995**, *67*, 279.
- (13) Ceperley, D. M. *Rev. Mod. Phys.* **1999**, *71*, S438.
- (14) Rabani, E.; Krilov, G.; Reichman, D. R.; Berne, B. J. *J. Chem. Phys.* **2005**, *123*, 84506.
- (15) Reichman, D. R.; Rabani, E. *Phys. Rev. Lett.* **2001**, *87*, 265702.
- (16) Rabani, E.; Reichman, D. R. *J. Chem. Phys.* **2002**, *116*, 6271.
- (17) Reichman, D. R.; Rabani, E. *J. Chem. Phys.* **2002**, *116*, 6279.
- (18) Rabani, E.; Reichman, D. R. *Europhys. Lett.* **2002**, *60*, 656.
- (19) Rabani, E.; Reichman, D. R.; Krilov, G.; Berne, B. J. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 1129.
- (20) Cao, J.; Voth, G. A. *J. Chem. Phys.* **1993**, *99*, 10070.
- (21) Cao, J.; Voth, G. A. *J. Chem. Phys.* **1994**, *100*, 5106.
- (22) Cao, J.; Martyna, G. J. *J. Chem. Phys.* **1996**, *104*, 2028.
- (23) Hone, T. D.; Voth, G. A. *J. Chem. Phys.* **2004**, *121*, 6412.
- (24) Hone, T. D.; Rossky, P. J.; Voth, G. A. *J. Chem. Phys.* **2006**, *124*, 154103.
- (25) Nakayama, A.; Makri, N. *J. Chem. Phys.* **2003**, *119*, 8592.
- (26) Nakayama, A.; Makri, N. *J. Chem. Phys.* **2006**, *125*, 024503.
- (27) Hone, T. D.; Poulsen, J. A.; Rossky, P. J.; Manolopoulos, D. E. *J. Phys. Chem. B* **2008**, *112*, 294.
- (28) Poulsen, J. A.; Nyman, G.; Rossky, P. J. *J. Chem. Theory Comput.* **2006**, *2*, 1482.
- (29) Poulsen, J. A.; Nyman, G.; Rossky, P. J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6709.
- (30) Miller III, T. H.; Manolopoulos, D. E. *J. Chem. Phys.* **2005**, *122*, 184503.
- (31) Miller III, T. H. *J. Chem. Phys.* **2008**, *129*, 194502.
- (32) Craig, I. R.; Manolopoulos, D. E. *J. Chem. Phys.* **2004**, *121*, 3368.
- (33) Craig, I. R.; Manolopoulos, D. E. *J. Chem. Phys.* **2005**, *122*, 084106.
- (34) Craig, I. R.; Manolopoulos, D. E. *J. Chem. Phys.* **2006**, *322*, 236.
- (35) Habershon, S.; Braams, B. J.; Manolopoulos, D. E. *J. Chem. Phys.* **2007**, *127*, 174108.
- (36) Rabani, E.; Reichman, D. R. *Annu. Rev. Phys. Chem.* **2005**, *56*, 157.
- (37) Kim, D.; Doll, J. D.; Gubernatis, J. E. *J. Chem. Phys.* **1997**, *106*, 1641.
- (38) Gallicchio, E.; Berne, B. J. *J. Chem. Phys.* **1994**, *101*, 9901.
- (39) Gallicchio, E.; Egorov, S. A.; Berne, B. J. *J. Chem. Phys.* **1998**, *109*, 7745.
- (40) Krilov, G.; Sim, E.; Berne, B. J. *J. Chem. Phys.* **2001**, *268*, 21.
- (41) Boninsegni, M.; Ceperley, D. M. *J. Low Temp. Phys.* **1996**, *104*, 339.
- (42) Tikhonov, A. N.; Goncharsky, A. V.; Stepanov, V. V.; Yagola, A. G. *Numerical Methods for the Solution of Ill-Posed Problems*; Kluwer Academic Publishers: London, 1995.
- (43) Hansen, P. H. *Rank-Deficient and Discrete Ill-Posed Problems*, SIAM: Philadelphia, PA, 1992.
- (44) Kowalczyk, P.; Tanaka, H.; Kanoh, H.; Kaneko, K. *Langmuir* **2004**, *20*, 2324.
- (45) von Szombathely, M.; Brauer, P.; Jaroniec, M. *Comput. Chem.* **1992**, *13*, 17.
- (46) Kowalczyk, P.; Terzyk, A. P.; Gauden, P. A. *Langmuir* **2002**, *18*, 5406.
- (47) Rubtsov, A. N.; Savkin, V. V.; Lichtenstein, A. I. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2005**, *72*, 035122.
- (48) Feynman, R. P.; Hibbs, A. *Quantum Mechanics and Path Integrals*; McGraw-Hill: New York, 1965.
- (49) Feynman, R. P. *Statistical Mechanics*; Benjamin: New York, 1972.
- (50) Aziz, R. A.; Slaman, M. J.; Koide, A.; Allnatt, A. R.; Meath, W. J. *Mol. Phys.* **1992**, *77*, 321.

JCTC

Journal of Chemical Theory and Computation

Gauge-Origin Independent Formulation and Implementation of Magneto-Optical Activity within Atomic-Orbital-Density Based Hartree–Fock and Kohn–Sham Response Theories

Thomas Kjærgaard* and Poul Jørgensen

The Lundbeck Foundation Center for Theoretical Chemistry, University of Aarhus, Langelandsgade 140, DK-8000 Århus C, Denmark

Andreas J. Thorvaldsen

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

Paweł Sałek

Department of Theoretical Chemistry, The Royal Institute of Technology SE-10691 Stockholm, Sweden

Sonia Coriani

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

Received April 6, 2009

Abstract: A Lagrangian approach has been used to derive gauge-origin independent expressions for two properties that rationalize magneto-optical activity, namely the Verdet constant $V(\omega)$ of the Faraday effect and the \mathcal{B} term of magnetic circular dichroism. The approach is expressed in terms of an atomic-orbital density-matrix based formulation of response theory and use London atomic orbitals to parametrize the magnetic field dependence. It yields a computational procedure which is both gauge-origin independent and suitable for linear-scaling at the level of time-dependent Hartree–Fock and density functional theory. The formulation includes a modified preconditioned conjugated gradient algorithm, which projects out the excited state component from the solution to the linear response equation. This is required when solving one of the response equations for the determination of the \mathcal{B} term and divergence is encountered if this component is not projected out. Illustrative results are reported for the Verdet constant of H_2 , HF, CO, N_2O , and $\text{CH}_3\text{CH}_2\text{CH}_3$ and for the \mathcal{B} term of pyrimidine, phosphabenzene, and pyridine. The results are benchmarked against gauge-origin independent CCSD values.

1. Introduction

It is well-known that a static magnetic field applied in the direction of propagation of linearly polarized light impinging

on matter induces an optical response, irrespective of the chirality of the sample. What is observed is either a rotation of the plane of polarization of the emerging light in the transparent regions of the sample (magneto-optical rotation, MOR, or magnetic circular birefringence MCB, also known as Faraday effect) or an ellipticity or dichroism in the

* To whom correspondence should be addressed. E-mail: tkjaergaard@chem.au.dk.

absorptive regions (magnetic circular dichroism, MCD). Both phenomena are collectively referred to as magneto-optical activity (MOA) and originate from a differential interaction of the sample with the right and left circularly polarized components of the plane-polarized light because of the presence of the perturbing magnetic field. For MOR the magnetic field causes the two components to propagate with different velocities, hereby, the rotation; for MCD the two components are absorbed to a different extent and the ellipticity occurs.^{1–6} Magnetic optical activity has turned out to become a valuable spectroscopic tool, in particular for studying the structure and electronic configurations of inorganic complexes, porphyrins, heme, and nonheme proteins.^{7–11}

MOR is conventionally rationalized in terms of the so-called Verdet constant,¹² which according to molecular perturbation theory, is connected to the dipole–dipole-magnetic dipole hyperpolarizability.^{3–5} The theoretical determination of the Verdet constant has been used as a probe of the electronic structure of the sample,^{13,14} to test the accuracy of new methods in computing high-order molecular properties,^{15–19} and even as source of information on structure–property relationships for organic molecules, which might be of importance in designing materials with outstanding magneto-optical properties.¹⁴

The induced dichroism is employed in the well-known MCD spectroscopy,^{20–23} which is widely used in inorganic chemistry, often with focus on the elucidation of the electronic structure of porphyrins and phthalocyanines,^{9,10} and in biological chemistry,^{11,24–26} for instance as a powerful probe of metalloenzyme active-sites.²⁴ MCD spectra are traditionally rationalized in terms of three magnetic rotatory strengths, known as the Faraday \mathcal{A} , \mathcal{B} , and \mathcal{C} terms.^{5,6,27–31} The \mathcal{A} term only contributes if either the ground or the excited state is degenerate, and the \mathcal{C} term only contributes if the ground state is degenerate. The \mathcal{B} term contributes irrespective of the degeneracy of the ground and final excited states and therefore describes the MCD spectrum of an electronic transition in a molecule without electronic state degeneracies. Even though it may be questioned whether it is necessary to separate the evaluation (and interpretation) of the MCD into \mathcal{A} and \mathcal{B} terms,³² the generally low symmetry of large organic molecules justifies the focus on the \mathcal{B} term as the \mathcal{A} and \mathcal{C} terms vanish in the absence of an axis of 3-fold or higher symmetry.^{4,22,23,33} After a few decades of relatively scarce attention, the computational determination of the MCD parameters and spectra is experiencing a renewed interest, both from the methodological and the applicative point of view, with a steady increase in the number of publications, in particular during the last five years.^{17,32,34–46}

A common difficulty in the calculation of magnetic properties is that approximate methods do not automatically guarantee gauge-origin independence of the results. The computed results might therefore end up depending on the origin of the vector potential. Only in the limit of a complete basis set do they become gauge-origin independent. A large basis set can then be expected to yield approximate

gauge invariance, but its use is not always a viable technique, in particular for very large systems.

For systems belonging to certain symmetry groups (e.g., atomic and centro-symmetric systems) the Verdet constant is known to be automatically origin independent, and it can be straightforwardly obtained from the dipole, dipole, magnetic dipole quadratic response function. Such an approach has been extensively used in the past at various levels of theory, from Hartree–Fock to coupled cluster and for various systems.^{14–16,47–49} For atoms, Cauchy expansions and linear response function approaches have also been employed.^{50,51}

The first gauge-origin independent approach to compute Verdet constants was proposed by Coriani et al.¹⁷ within coupled cluster singles and doubles (CCSD) response theory and made use of London atomic orbitals (LAO),^{52–54} which introduce local gauge-origins to define the vector potential of the external magnetic field. A similar strategy was adopted a few years later by Banerjee and co-workers^{18,19} at the time-dependent (TD) density functional theory (DFT) level, in a molecular orbital-based formulation.

Concerning the MCD strengths, early methods to calculate the \mathcal{B} term have in general been based on sum-overstates (SOS) procedures,^{55–61} often at a semiempirical level. Seth et al.⁴¹ presented a \mathcal{B} -SOS method at the TD-DFT level of theory in a molecular orbital based formulation. Any SOS method suffers from errors caused by truncations in the number of excitations that are considered.

The SOS method is gauge-origin dependent, like any numerical calculation which is forced to use an incomplete basis set of field-free eigenfunctions.^{33,62} Caldwell and Eyring^{4,62} have shown that the error introduced by using a limited basis set is minimized at the center of the charge density, and several authors have reported that the location of the origin within the framework of the molecule yields qualitatively correct results.^{63,64}

In ref 34 the \mathcal{B} term was evaluated, for naturally gauge-origin independent molecules,³³ as the first residue of the frequency dependent quadratic response functions.³⁴ The method can be considered superior to the SOS procedure, but no general solution for the treatment of the gauge-origin problem was proposed.³⁴ Note that the usual interpretation of MCD in terms of magnetic mixing is lost in such an approach, as reference to intermediate states is avoided. An extension of the approach to include solvent effects was proposed in ref 37.

Several papers have recently appeared on the computational simulation of MCD spectra,^{17,32,35–46,65} including \mathcal{A} , \mathcal{B} , and \mathcal{C} terms. We bring attention in particular to the novel complex polarization approach of ref 38 and the rather similar approach based on the imaginary part of the Verdet constant using damped time-dependent density functional theory in ref 44. In the latter, in particular, the use of London orbitals in the determination of the damped Verdet constant allows to remove the gauge-origin dependence problem at the DFT level. Ref 42 applies a magnetically perturbed TD-DFT⁶⁶ to calculate the \mathcal{B} term of MCD and describes both a SOS approach and a direct approach. The direct approach is similar to the approach presented by Coriani et al.³⁴

The use of London orbitals to remove the gauge-origin dependence in the calculation of the \mathcal{B} term was first proposed in 1972 by Seaman and Linderberg⁶⁷ and implemented within a finite perturbation (FP) method. Coriani et al. presented in ref.,¹⁷ together with the gauge-independent expression for the Verdet constant, also a gauge-origin independent formulation of the \mathcal{B} term at the CCSD level using LAOs. The method was later applied to a number of selected molecules³⁶ and gave \mathcal{B} terms of a high quality, but it is hampered by its computational cost and is limited to molecules of the size of pyrimidine and phosphabenzene.

The Hartree–Fock approximation may be used to calculate \mathcal{B} term of larger molecules, but early studies showed that the Hartree–Fock method may not even produce the correct sign for the \mathcal{B} term.^{34,37} It was also concluded that electron correlation effects are, as for other optical properties, important for determining the \mathcal{B} term.

One way to efficiently include electron correlation effects for large molecules is via density functional theory (DFT). DFT is known to give reasonable accuracy at a low computational cost,^{68–70} and recent linear scaling implementations make Kohn–Sham DFT (along with Hartree–Fock) applicable to molecules consisting of more than 1000 atoms. In ref 71, for instance, linear scaling of the DFT Kohn–Sham (KS) method has been obtained by a reformulation in a orthogonal atomic orbital (OAO) basis, the Löwdin basis. The orthogonal OAO basis is advantageous to use to describe local phenomena such as the Coulomb cusp. In addition to exploiting locality the Löwdin orthogonalization reduces the condition number of the KS matrix, a very attractive feature when solving the response equations of DFT.

The locality of the orbitals makes matrices sparse, and with the use of sparse matrix algebra,⁷² it is possible to avoid the high scaling of the standard electronic structure models and achieve linear scaling. The use of the method requires an efficient and linear-scaling transformation from the AO basis to the Löwdin OAO basis, which has been developed and implemented by Jansík et al.⁷³

Coriani et al.⁷⁴ addressed the problem of the computation of molecular properties at a linear cost and used the Löwdin OAO basis to obtain an atomic orbital based response solver which allows for the linear scaling calculation of excitation energies, transition strengths and frequency dependent linear response functions (polarizabilities) at the Hartree–Fock and Kohn–Sham (TD-DFT) levels of theory. The method has been recently extended to the general frequency-dependent quadratic response function and its residues by Kjergaard et al.⁷⁵

We here present a gauge-origin independent formulation of the Verdet constant and of the \mathcal{B} terms of MCD within the linear scaling framework of Hartree–Fock/Kohn–Sham-DFT response theory of ref 74. Gauge-origin independent analytic expressions for the two MOA properties are found using a Lagrangian technique, where each property is expressed as a total derivative, with respect to the strength of the external magnetic field, of an appropriate linear-response functional whose magnetic field dependence is parametrized by means of LAOs. By choosing, as starting expressions, lower-order property AO-based expressions,^{74,76}

we automatically obtain working equations which are prone to linear scaling for sufficiently sparse matrices. The atomic orbital basis, in addition to guaranteeing scalability and ease of parallelization, also represents a “natural” framework for deriving property expressions for perturbation dependent basis sets.

The procedure adopted here is presented in more general terms in ref 77, where the focus lies on geometrical derivatives. It also bears similarities with Furche and co-workers^{78,79} conventional molecular-orbital formulation of excited state gradients and vibrational Raman intensities using time-dependent density functional theory. Geometric and magnetic perturbations can in fact be treated on the same footing, as involving different types of perturbation-dependent basis sets. Similarities may also be found with the method presented by Thorvaldsen et al.⁸⁰ for the calculation at the Kohn–Sham level of molecular properties to arbitrary order, in which the quasienergy and Lagrangian formalisms are combined to derive response functions by differentiation of the quasienergy derivative Lagrangian using the elements of the AO density matrix as variational parameters, and which was implemented, at the Hartree–Fock level only, to obtain for instance gauge-origin independent values of the mixed electric and magnetic first hyperpolarizability entering the temperature-independent term of Buckingham’s birefringence.⁸¹ Such quantity is somewhat connected to the Verdet constant, as discussed for instance in refs 51 and 82.

When London orbitals are not used we recover, as a subset of our derivation, the standard expression for the quadratic response function $\langle\langle\hat{\mu}_\alpha; \hat{\mu}_\beta, \hat{m}_\gamma\rangle\rangle_{\omega,0}$ and its residues, with $\hat{\mu}$ being the electric dipole operator and \hat{m} the magnetic dipole, that is, the standard quadratic response expressions for the Verdet constant $V(\omega)$ and the \mathcal{B} term of MCD.³⁴ Note that when evaluating the \mathcal{B} term, we also ensure that unphysical divergencies related to singularities in the response equations are projected out.

The method is mainly benchmarked against new and previous CCSD results.^{17,36} This is expected to be of particular relevance for MCD, as it will increase the reliability of these calculations for the assignment of excited states for large molecules. The CCSD study of Kjergaard et al.³⁶ showed that Kohn–Sham DFT calculations cannot directly be benchmarked against experimental results because of the cancellation between positive and negative contributions.

The paper is organized as follows. In the Theory section, we define first the key quantities which constitute the formal background of our derivation. We then present the Lagrange method for the properties of interest. In the Results and Discussion section, we report pilot numerical results for a few selected cases, addressing both gauge-origin independence and convergence issues.

2. Theory

2.1. Definition of the MOA Properties: Derivative Expressions of $V(\omega)$ and $\mathcal{B}(n \rightarrow j)$. Following the approach of Coriani et al.,¹⁷ we define the Verdet constant $V(\omega)$ of the Faraday effect as⁵

$$V(\omega) = \left\langle \left(\frac{d\phi}{dB_z} \right)_{B=0} \right\rangle = \omega C \epsilon_{\alpha\beta\gamma} \left(\frac{d\alpha'_{\alpha\beta}(-\omega; \omega)}{dB_\gamma} \right)_{B=0} = C \omega \epsilon_{\alpha\beta\gamma} \alpha_{\alpha,\beta,\gamma}^{(m)} \quad (1)$$

where ϕ is the optical rotation, $\alpha'_{\alpha\beta}(-\omega; \omega)$ is the (magnetic-field perturbed) antisymmetric electric dipole polarizability (the imaginary part of the complex polarizability $\tilde{\alpha}_{\alpha\beta}(-\omega; \omega)$), ω is the frequency of the incident polarized light and $\epsilon_{\alpha\beta\gamma}$ is the Levi–Civita tensor.³ C is a collection of fundamental constants, $C = (1)/(12)\mu_0 c N$, where c is the speed of light in vacuo, μ_0 is the permeability of free space, and N is the number density. B_z is the z Cartesian component of the magnetic field induction. Note that this expression only holds for an isotropic sample of closed shell molecules, otherwise an additional temperature dependent term would appear.³

Similarly, we identify the Faraday $\mathcal{B}(n \rightarrow j)$ from the derivative of a (magnetic-field perturbed) one-photon dipole transition strength $S_{nj}^{\alpha\beta} = \langle n | \hat{\mu}_{\alpha j} \rangle \langle j | \hat{\mu}_{\beta n} \rangle \equiv \mu_{\alpha}^{nj} \mu_{\beta}^{jn}$

$$\mathcal{B}(n \rightarrow j) = \frac{1}{2} \epsilon_{\alpha\beta\gamma} \mathcal{F} \left(\frac{dS_{nj}^{\alpha\beta}}{dB_\gamma} \right)_{B=0} \quad (2)$$

where $|n\rangle$ is the ground state.^{17,67}

We mention here that a derivative expression was recently proposed by Seth and co-workers³⁹ also for the \mathbf{A} term

$$\mathcal{A}(n \rightarrow j) = -\frac{1}{2} \epsilon_{\alpha\beta\gamma} \sum_r \left(\frac{\partial \omega_r}{\partial B_\gamma} \right)_{B=0} \mathcal{F}(\mu_{\alpha}^{nj_r} \mu_{\beta}^{r_n}) \quad (3)$$

where ω_r is the excitation frequency from the ground state to the degenerate excited state $|j_r\rangle$, with r running on the number of degenerate states. We defer the calculation of the \mathcal{A} term to a later publication.

According to the definitions above, the two MOA properties are hence magnetic-field derivatives of quantities that can be obtained, within response theory,⁸³ from the so-called linear response function. It is well-known that the electric dipole polarizability corresponds to the linear response function

$$\alpha_{\alpha\beta}(-\omega; \omega) = -\langle \langle \hat{\mu}_{\alpha}; \hat{\mu}_{\beta} \rangle \rangle_{\omega} \quad (4)$$

where μ_{α} is the α Cartesian component of the electric dipole operator. Similarly, the transition strength is obtained from the residue of the linear response function

$$S_{nj}^{\alpha\beta} = \lim_{\omega \rightarrow \omega_j} (\omega - \omega_j) \langle \langle \hat{\mu}_{\alpha}; \hat{\mu}_{\beta} \rangle \rangle_{\omega} \quad (5)$$

The definitions above allow us to derive gauge-origin independent expressions by setting up a two-step approach that starts from the analytic expressions for the linear response functions in an AO based framework (section 2.2) and where the magnetic field dependence in the linear response quantities is parametrized through the LAOs (section 2.3). By exploiting a Lagrangian approach, we can moreover take advantage of the $2n + 1$ and $2n + 2$ rules to avoid computing the responses to the magnetic field of the linear response vectors and/or excitation vectors, and of the

Table 1. Explicit Expressions for the Various Vectors, Matrices and Super-Matrices Entering the Linear Response Function in the Hartree-Fock Case^a

$E_m^{[1]}$	$= \text{Tr} \mathbf{O}_m^{\dagger} [\mathbf{F}, \mathbf{S}]_{\mathbf{D}}$
$\mathbf{B}_m^{[1]}$	$= \text{Tr} \mathbf{O}_m^{\dagger} [\mathbf{B}, \mathbf{S}]_{\mathbf{D}}$
$\mathbf{A}_m^{[1]}$	$= \text{Tr} \mathbf{O}_m^{\dagger} [\mathbf{A}, \mathbf{S}]_{\mathbf{D}}$
$E_{mh}^{[2]}$	$= \text{Tr} \mathbf{F} [[\mathbf{O}_m, \mathbf{D}]_s, \mathbf{O}_m^{\dagger}]_s + \text{Tr} \mathbf{G} ([\mathbf{O}_m, \mathbf{D}]_s) [\mathbf{D}, \mathbf{O}_m^{\dagger}]_s$
$\mathbf{S}_{mh}^{[2]}$	$= \text{Tr} \mathbf{O}_m^{\dagger} \mathbf{S} [\mathbf{D}, \mathbf{O}_m]_s \mathbf{S}$
$E_{mj}^{[3]}$	$= \text{Tr} \mathbf{F} [[\mathbf{O}_m, [\mathbf{O}_j, \mathbf{D}]_s], \mathbf{O}_m^{\dagger}]_s + \text{Tr} \mathbf{G} ([\mathbf{O}_j, \mathbf{D}]_s) [[\mathbf{O}_m, \mathbf{D}]_s, \mathbf{O}_m^{\dagger}]_s + \text{Tr} \mathbf{G} ([\mathbf{O}_m, \mathbf{D}]_s) [[\mathbf{O}_j, \mathbf{D}]_s, \mathbf{O}_m^{\dagger}]_s + \text{Tr} \mathbf{G} ([\mathbf{O}_m, [\mathbf{O}_j, \mathbf{D}]_s]) [\mathbf{D}, \mathbf{O}_m^{\dagger}]_s$
$\mathbf{S}_{mj}^{[3]}$	$= -\text{Tr} \mathbf{O}_m^{\dagger} \mathbf{S} [\mathbf{O}_m, [\mathbf{D}, \mathbf{O}_j]_s] \mathbf{S}$
$\mathbf{B}_{mj}^{[2]}$	$= -\text{Tr} \mathbf{B} [[\mathbf{O}_j, \mathbf{D}]_s, \mathbf{O}_m^{\dagger}]_s$
$\mathbf{A}_{mj}^{[2]}$	$= \text{Tr} \mathbf{A}^{\dagger} [[\mathbf{O}_j, \mathbf{D}]_s, \mathbf{O}_m^{\dagger}]_s$

^a The superscript refers to the order of differentiation with respect to the variational parameters \mathbf{X} in the density matrix. Note that $(\partial \mathbf{D}(\mathbf{X})) / (\partial X_i) = -[\mathbf{O}_i, \mathbf{D}]_s$.

variational parameters involved in the wavefunction/density optimization (section 2.4).

2.2. AO-Based Linear Response Theory. The linear response function is the first-order term in the perturbative expansion, with respect to a (periodic) perturbation, $\hat{V}^{\omega} = \int_{-\infty}^{+\infty} \hat{V}^{\omega} \exp(-i\omega t) d\omega$, of the time-dependent expectation value of an observable represented by the time-independent operator \hat{A}

$$\langle \hat{A} \rangle(t) = \langle \hat{A} \rangle_0 + \int \langle \langle \hat{A}; \hat{V}^{\omega} \rangle \rangle_{\omega} \exp(-i\omega t) d\omega + \dots \quad (6)$$

where $\langle \langle \hat{A}; \hat{V}^{\omega} \rangle \rangle_{\omega}$ is the linear response function in the frequency domain. The linear response function (LRF) is required for the calculation of the Verdet constant, according to eqs 4 and 1.

Assuming implicit summation over repeated indices and introducing the symbol \hat{B} in place of \hat{V}^{ω} , the linear response function (LRF) expression according to the AO parametrization of response theory of refs 74 and 76 is

$$\langle \langle \hat{A}; \hat{B} \rangle \rangle_{\omega} = -\mathbf{A}^{[1]\dagger} \mathbf{b}^{\omega} \equiv -\text{Tr} \{ \mathbf{A}^{[1]\dagger} \mathbf{b}^{\omega} \} \quad (7)$$

where the elements b_m^{ω} of the response vector \mathbf{b}^{ω} are obtained from the solution of the linear response equation

$$(\mathbf{E}^{[2]} - \omega \mathbf{S}^{[2]}) \mathbf{b}^{\omega} = \mathbf{B}^{[1]} \quad (8)$$

Above, $\mathbf{A}^{[1]}$ is the so-called *property gradient* relative to the \hat{A} operator, whereas $\mathbf{B}^{[1]}$ is the *property gradient* relative to the external perturbation described by V^{ω} ($\equiv \hat{B}$) and $\mathbf{E}^{[2]}$ and $\mathbf{S}^{[2]}$ are the generalized (RPA) electronic Hessian and metric matrices in the AO basis.⁷⁶ Their explicit expressions are collected in Table 1. Note that the elements of $\mathbf{E}^{[2]}$ are here defined with sign opposite compared to the one given in eq 65 of ref 76. The reasons for such a sign change are explained in detail in ref 75.

Note also the use of italic boldface characters to indicate vectors and matrices in a so-called *supermatrix* notation,^{83,84} and of roman bold face characters to indicate true matrices in the AO space (of dimension N_{AO}^2). According to the supermatrix notation, an AO matrix element, for instance the $B_{\mu\nu}^{[1]}$ element of the property gradient matrix $\mathbf{B}^{[1]}$, corresponds to the element $B_m^{[1]}$ of a property gradient (*column*) vector $\mathbf{B}^{[1]}$, where m is ordered such that the

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

$$M_j = \text{Tr} \mathbf{O}_j^\dagger \mathbf{M} = \text{Tr} \mathbf{O}_j \mathbf{M}^\dagger \quad (9)$$

which associates the j element of vector \mathbf{M} with the $M_{\mu\nu}$ element of the matrix \mathbf{M} in the AO basis. The operators \mathbf{O}_j^\dagger and \mathbf{O}_j are defined in ref 76 to be

$$\mathbf{O}_j = \begin{cases} \mathbf{E}_{\mu\nu} & j > 0 \\ \mathbf{E}_{\nu\mu} & j < 0 \end{cases} \quad (10)$$

where $\mathbf{E}_{\mu\nu}$ is a unit matrix with elements given by

$$[\mathbf{E}_{\mu\nu}]_{\rho\sigma} = \delta_{\mu\rho} \delta_{\nu\sigma} \quad (11)$$

We also introduce the expansions

$$\mathbf{b}^\omega = \sum_m b_m^\omega \mathbf{O}_m; \quad \mathbf{b}^{\omega\dagger} = \sum_m b_m^{\omega\dagger} \mathbf{O}_m^\dagger \quad (12)$$

where the first vector is a column vector, and the second one is a row vector. In the AO basis their matrix representations will be indicated as \mathbf{b}^ω and $\mathbf{b}^{\omega\dagger}$, respectively.

As indicated in Table 1, the property gradients and generalized Hessian and metric contain matrix products of integral matrices in the AO space, namely the AO overlap matrix \mathbf{S} , the AO integral matrices \mathbf{A} and \mathbf{B} for the \hat{A} and \hat{B} operators, respectively, the AO density matrix \mathbf{D} , the Fock

$$\mathbf{F} = \mathbf{h} + \mathbf{G}^{\text{HF}}(\mathbf{D}) \quad (13)$$

or Kohn–Sham⁷⁵

$$\mathbf{F} = \mathbf{h} + \mathbf{G}^{\text{HF}}(\mathbf{D}) + \mathbf{F}^{\text{xc}} \quad (14)$$

matrices, where

$$\mathbf{G}_{\mu\nu}^{\text{HF}}(\mathbf{D}) = \sum_{\rho\sigma} D_{\rho\sigma} [\zeta_{\mu\nu\rho\sigma} - w_x \zeta_{\mu\sigma\rho\nu}] \quad (15)$$

The scaling factor w_x is equal to 1 for Hartree–Fock; in the DFT case it is only nonzero for hybrid theories. The last term in eq 14 is the derivative of the exchange–correlation functional $E_{\text{xc}}[\rho]$, which depends on the density ρ

$$F_{\mu\nu}^{\text{xc}} = \frac{\partial E_{\text{xc}}[\rho]}{\partial D_{\nu\mu}} \quad (16)$$

Expressing the density ρ in the AO basis as

$$\rho(\mathbf{r}) = \sum_{\mu\nu} \chi_\mu^*(\mathbf{r}) \chi_\nu(\mathbf{r}) D_{\nu\mu} = \sum_{\mu\nu} \Omega_{\mu\nu}(\mathbf{r}) D_{\nu\mu} \quad (17)$$

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

$$v_{\text{xc}}(\mathbf{r}) = \frac{\delta E_{\text{xc}}[\rho]}{\delta \rho(\mathbf{r})} \quad (18)$$

it is seen that

$$F_{\mu\nu}^{\text{xc}} = \int \frac{\delta E_{\text{xc}}[\rho]}{\delta \rho(\mathbf{r})} \frac{\partial \rho(\mathbf{r})}{\partial D_{\nu\mu}} d\mathbf{r} = \int v_{\text{xc}}(\mathbf{r}) \Omega_{\mu\nu}(\mathbf{r}) d\mathbf{r} \quad (19)$$

is the AO matrix representation of the exchange–correlation potential. Within the exponential parametrization of the density matrix, density matrices are related through the transformation

$$\mathbf{D}(\mathbf{X}) = \exp(-\mathbf{X}\mathbf{S})\mathbf{D} \exp(\mathbf{S}\mathbf{X}) = \mathbf{D} + [\mathbf{D}, \mathbf{X}]_{\mathbf{S}} + \frac{1}{2}[[\mathbf{D}, \mathbf{X}]_{\mathbf{S}}, \mathbf{X}]_{\mathbf{S}} + \dots \quad (20)$$

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

$$\mathbf{X} = \mathcal{P}(\mathbf{X}) \equiv \mathbf{P}_o \mathbf{X} \mathbf{P}_v^\dagger + \mathbf{P}_v \mathbf{X} \mathbf{P}_o^\dagger \quad (21)$$

\mathbf{P}_o and \mathbf{P}_v are projectors onto the occupied and virtual orbital spaces, respectively

$$\mathbf{P}_o = \mathbf{D}\mathbf{S} \quad (22)$$

$$\mathbf{P}_v = \mathbf{I} - \mathbf{D}\mathbf{S} \quad (23)$$

fulfilling the idempotency ($\mathbf{P}_o^2 = \mathbf{P}_o$ and $\mathbf{P}_v^2 = \mathbf{P}_v$) and orthogonality relations ($\mathbf{P}_o \mathbf{P}_v = \mathbf{P}_v \mathbf{P}_o = 0$ and $\mathbf{P}_o^\dagger \mathbf{S} \mathbf{P}_v = \mathbf{P}_v^\dagger \mathbf{S} \mathbf{P}_o = 0$). We have used the so-called S-commutator

$$[\mathbf{D}, \mathbf{X}]_{\mathbf{S}} = \mathbf{D}\mathbf{S}\mathbf{X} - \mathbf{X}\mathbf{S}\mathbf{D} \quad (24)$$

For later convenience we also introduce a generalized M commutator as

$$[\mathbf{L}, \mathbf{N}]_{\mathbf{M}} = \mathbf{L}\mathbf{M}\mathbf{N} - \mathbf{N}\mathbf{M}\mathbf{L} \quad (25)$$

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

$$(\mathbf{E}^{[2]} - \omega_f \mathbf{S}^{[2]}) \mathbf{b}^f = 0 \quad (26)$$

The excited state vector is normalized over the generalized metric matrix $\mathbf{S}^{[2]}$, that is

$$\mathbf{b}^{f\dagger} \mathbf{S}^{[2]} \mathbf{b}^f = 1 \quad (27)$$

We note that $\mathbf{S}^{[2]}$ is *not* positive definite.

To calculate the Faraday \mathcal{B} term, the residue of the linear response function is required (see eqs 2 and 5). The expression for the residue of the linear response function is, according to the AO parametrization of response theory, given by⁷⁴

$$\lim_{\omega \rightarrow \omega_f} (\omega - \omega_f) \langle \langle A; B \rangle \rangle_\omega = (\mathbf{A}^{[1]\dagger} \mathbf{b}^f) (\mathbf{b}^{f\dagger} \mathbf{B}^{[1]}) = \text{Tr}\{\mathbf{A}^{[1]\dagger} \mathbf{b}^f\} \text{Tr}\{\mathbf{b}^{f\dagger} \mathbf{B}^{[1]}\} \quad (28)$$

2.3. Parametrization of the Magnetic Field Dependence: The LAO Basis and the (Differentiated) Integral Matrices in the LAO Basis. London atomic orbitals are used to parametrize the magnetic field dependence of the properties

to be differentiated. For a given atomic orbital $\chi(\mathbf{r}_M)$ centered on the nucleus M at position \mathbf{R}_M , the London orbital⁵² is defined

$$\omega_\mu(\mathbf{r}_M, \mathbf{B}) = \exp\left\{-\frac{i}{2}\mathbf{B} \cdot [(\mathbf{R}_M - \mathbf{O}) \times \mathbf{r}]\right\}\chi_\mu(\mathbf{r}_M) \quad (29)$$

where $\chi_\mu(\mathbf{r}_M)$ is a standard atomic orbital. Hence, the LAO depends parametrically on the magnetic field \mathbf{B} , the gauge origin \mathbf{O} and the orbital position \mathbf{R}_M .

The LAOs introduce a magnetic field dependence in the integral matrices through exponential factors modifying the operators

$$O_{\mu\nu}(\mathbf{B}) = \langle \mu | \exp\left(\frac{i}{2}\mathbf{B} \cdot \mathbf{R}_{MN} \times \mathbf{r}\right) \hat{O} | \nu \rangle \quad (30)$$

where \hat{O} is any one-electron operator or a Cartesian component operator, and

$$g_{\mu\nu\rho\sigma}(\mathbf{B}) = \langle \mu\nu | \exp\left\{\frac{i}{2}\mathbf{B} \cdot (\mathbf{R}_{MN} \times \mathbf{r}_1 + \mathbf{R}_{RS} \times \mathbf{r}_2)\right\} r_{12}^{-1} | \rho\sigma \rangle \quad (31)$$

are the magnetic-field dependent two-electron integrals; $\mathbf{R}_{MN} = \mathbf{R}_M - \mathbf{R}_N$ is the distance vector between nucleus M and nucleus N . Note that $\chi_\nu(\mathbf{r}_N)$, $\chi_\rho(\mathbf{r}_R)$, and $\chi_\sigma(\mathbf{r}_S)$ are centered at \mathbf{R}_N , \mathbf{R}_R , and \mathbf{R}_S , respectively. The one-electron operators of relevance in our specific case are the identity operator (for the overlap matrix integrals), the electric dipole operator ($\hat{O} = \mu_x, \mu_y, \mu_z$), and the one-electron Hamiltonian h_N

$$h_N(\mathbf{r}; \mathbf{B}, \mathbf{m}) = \frac{1}{2}\pi_N^2 - \sum_k \frac{Z_k}{|\mathbf{r} - \mathbf{R}_k|} \quad (32)$$

$$\pi_N = -i\nabla + \frac{1}{2}\mathbf{B} \times \mathbf{r}_N + \alpha^2 \sum_K \frac{\mathbf{m}_K \times \mathbf{r}_K}{r_K^3} \quad (33)$$

The second term in the kinetic momentum operator π_N represents an externally applied uniform magnetic field \mathbf{B} , and the third term represents the field from a nuclear point magnetic moment \mathbf{m}_K , where α is the fine structure constant. The integrals over LAOs have been shown to be origin independent.⁵³

Explicit expressions for the differentiated overlap, electric dipole, and Hamilton operator matrices (including the electron repulsion) have previously been reported,^{53,54} and we refer the interested reader to the original papers. In the DFT case, however, we need to consider the modifications resulting from the use of LAOs in the exchange-correlation contribution \mathbf{F}^{xc} to the Kohn–Sham matrix in eq 14 and in the exchange correlation contribution $\mathbf{G}^{\text{xc}}(\mathbf{M})$ to the (generalized) Kohn–Sham Hessian $\mathbf{E}^{[2]}$ (see Table 1 and refs 75 and 76), and the resulting magnetic field derivatives. The evaluation of these terms is described in details in Appendix B.

For completeness, we report here the final result, which has been derived under the assumption that the exchange correlation energy is the integral over all space of some functional $f = f[\rho, \xi]$, which depends on the density ρ and on the norm of the density gradient $\xi = |\nabla\rho|$. The exchange correlation contribution to the Kohn–Sham matrix is

$$F_{\mu\nu}^{\text{xc}} = \int \frac{\partial f}{\partial \rho} \Omega_{\mu\nu}(\mathbf{r}) d\mathbf{r} + \int \frac{\partial f}{\partial \xi} \frac{\partial \xi}{\partial \nabla \rho} \nabla \Omega_{\mu\nu}(\mathbf{r}) d\mathbf{r} \quad (34)$$

and its magnetic field derivative (see ref 85) becomes

$$\left. \frac{\partial F_{\mu\nu}^{\text{xc}}}{\partial \mathbf{B}} \right|_{\mathbf{B}=0} = \frac{i}{2} \int (\mathbf{R}_{MN} \times \mathbf{r}) \left[\frac{\partial f}{\partial \rho(\mathbf{r})} \Omega_{\mu\nu}(\mathbf{r}) + \frac{\partial f}{\partial \xi} \frac{\nabla \rho(\mathbf{r})}{\xi} \nabla \Omega_{\mu\nu}(\mathbf{r}) \right] d\mathbf{r} + \frac{i}{2} \int (\mathbf{R}_{MN} \times \frac{\nabla \rho(\mathbf{r})}{\xi}) \frac{\partial f}{\partial \xi} \Omega_{\mu\nu}(\mathbf{r}) d\mathbf{r} \quad (35)$$

For the exchange-correlation contribution to the (generalized) Kohn–Sham Hessian, we obtain

$$G_{\mu\nu}^{\text{xc}}(\mathbf{M}) = \sum_{\rho\sigma} M_{\sigma\rho} \int \left[\frac{\partial^2 f}{\partial \rho^2} \Omega_{\mu\nu} \Omega_{\rho\sigma} + \frac{\partial^2 f}{\partial \rho \partial \xi} \left(\Omega_{\mu\nu} \frac{\nabla \rho}{\xi} \nabla \Omega_{\rho\sigma} + \Omega_{\rho\sigma} \frac{\nabla \rho}{\xi} \nabla \Omega_{\mu\nu} \right) \right] d\mathbf{r} + \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\partial^2 f}{\partial \xi^2} \left(\frac{\nabla \rho}{\xi} \nabla \Omega_{\mu\nu} \right) \left(\frac{\nabla \rho}{\xi} \nabla \Omega_{\rho\sigma} \right) d\mathbf{r} \quad (36)$$

where \mathbf{M} is a general matrix, which in our case corresponds either to the \mathbf{D}^b matrix given in eq 48 or to the \mathbf{D}^f matrix given in eq 68. Its magnetic field derivative becomes

$$\begin{aligned} \frac{\partial G_{\mu\nu}^{\text{xc}}}{\partial \mathbf{B}} &= \sum_{\rho\sigma} M_{\sigma\rho} \frac{i}{2} \int (\mathbf{R}_{MN} \times \mathbf{r}) \left(\frac{\partial^2 f}{\partial \rho^2} \Omega_{\mu\nu} \Omega_{\rho\sigma} + \frac{\partial^2 f}{\partial \rho \partial \xi} \Omega_{\mu\nu} \left(\frac{\nabla \rho}{\xi} \nabla \Omega_{\rho\sigma} \right) + \frac{\partial^2 f}{\partial \rho \partial \xi} \left(\frac{\nabla \rho}{\xi} \nabla \Omega_{\mu\nu} \right) \Omega_{\rho\sigma} + \frac{\partial^2 f}{\partial \xi^2} \left(\frac{\nabla \rho}{\xi} \nabla \Omega_{\mu\nu} \right) \left(\frac{\nabla \rho}{\xi} \nabla \Omega_{\rho\sigma} \right) \right) d\mathbf{r} + \sum_{\rho\sigma} M_{\sigma\rho} \frac{i}{2} \times \\ &\int (\mathbf{R}_{RS} \times \mathbf{r}) \left(\frac{\partial^2 f}{\partial \rho^2} \Omega_{\mu\nu} \Omega_{\rho\sigma} + \frac{\partial^2 f}{\partial \rho \partial \xi} \Omega_{\rho\sigma} \left(\frac{\nabla \rho}{\xi} \nabla \Omega_{\mu\nu} \right) + \frac{\partial^2 f}{\partial \rho \partial \xi} \left(\frac{\nabla \rho}{\xi} \nabla \Omega_{\rho\sigma} \right) \Omega_{\mu\nu} + \frac{\partial^2 f}{\partial \xi^2} \left(\frac{\nabla \rho}{\xi} \nabla \Omega_{\mu\nu} \right) \left(\frac{\nabla \rho}{\xi} \nabla \Omega_{\rho\sigma} \right) \right) d\mathbf{r} + \\ &\sum_{\rho\sigma} M_{\sigma\rho} \frac{i}{2} \int \left(\mathbf{R}_{MN} \times \nabla \rho(\mathbf{r}) \right) \left(\frac{\partial^2 f}{\partial \rho \partial \xi} \frac{1}{\xi} \Omega_{\mu\nu} \Omega_{\rho\sigma} + \frac{\partial^2 f}{\partial \xi \partial \xi} \frac{1}{\xi^2} \Omega_{\mu\nu} (\nabla \rho \nabla \Omega_{\rho\sigma}) \right) d\mathbf{r} + \sum_{\rho\sigma} M_{\sigma\rho} \frac{i}{2} \int (\mathbf{R}_{RS} \times \nabla \rho(\mathbf{r})) \times \\ &\left(\frac{\partial^2 f}{\partial \rho \partial \xi} \frac{1}{\xi} \Omega_{\mu\nu} \Omega_{\rho\sigma} + \frac{\partial^2 f}{\partial \xi^2} \frac{1}{\xi^2} \Omega_{\rho\sigma} (\nabla \rho \nabla \Omega_{\mu\nu}) \right) d\mathbf{r} + G_{\mu\nu}^{\text{xc}} \left(\frac{\partial \mathbf{M}}{\partial \mathbf{B}} \right) \end{aligned} \quad (37)$$

The last term is straightforwardly obtained and is therefore not explicitly given. For the details on the derivation of the above equations, see Appendix B.

2.4. Lagrangian Approach. Having introduced all key ingredients, we can now exploit the Lagrangian approach to derive efficient computational expressions for the two quantities of interest. We will illustrate the procedure for each property separately.

2.4.1. Lagrangian Formulation of the Verdet Constant. To derive a computational expression of the Verdet constant, the Lagrangian functional is built

$$\mathcal{L}^V = \mathbf{A}^{[1]\dagger} \mathbf{b}^\omega - \bar{\lambda}^\dagger [(\mathbf{E}^{[2]} - \omega \mathbf{S}^{[2]}) \mathbf{b}^\omega - \mathbf{B}^{[1]}] - \bar{\mathbf{X}}^\dagger \mathbf{E}^{[1]} \quad (38)$$

from the linear response function in the AO basis,⁷⁴ eq 7, plus two sets of constraint equations, each multiplied with appropriate Lagrange multipliers. The constraint equations are the equations for the linear response vectors, eq 8, and the optimization condition for the “orbital” parameters \mathbf{X} in the exponential parametrization of the density matrix.

$$\mathbf{E}^{[1]} = \mathbf{FDS} - \mathbf{SDF} = 0 \quad (39)$$

$\bar{\lambda}^\dagger$ and $\bar{\mathbf{X}}^\dagger$ are referred to as the property and orbital Lagrange multipliers, respectively.

Imposing that the functional is variational with respect to all parameters, that is,

$$\frac{\partial \mathcal{L}^V}{\partial \zeta_i} = 0; \quad \forall \zeta_i \in \{b_i^\omega, \bar{\lambda}_i, \bar{X}_i, X_i\} \quad (40)$$

yields the response equations that need to be solved, in accordance to the $2n + 1$ and $2n + 2$ rules. Differentiation with respect to all $\bar{\lambda}_i$ returns the response equation for the response vectors $\mathbf{b}^{\omega\omega}$ (eq 8); differentiation with respect to all \bar{X}_i returns the orbital optimization condition (eq 39); differentiation with respect to b_i^ω gives the response equation that determines the multipliers $\bar{\lambda}_i$, which is in the form of an adjoint response equation

$$[(\mathbf{E}^{[2]} - \omega \mathbf{S}^{[2]}) \bar{\lambda}]^\dagger = [\mathbf{A}^{[1]}]^\dagger \quad (41)$$

We may therefore in the following write $\bar{\lambda} = \mathbf{a}^\omega$ to stress the similarity between $\bar{\lambda}$ and \mathbf{b}^ω . Differentiation of \mathcal{L}^V with respect to the (nonredundant) orbital parameters X_i gives the response equation to determine the multipliers $\bar{\mathbf{X}}$, which can be recast in the form

$$\mathbf{E}^{[2]} \bar{\mathbf{X}} = \eta \quad (42)$$

where the right-hand-side vector η is given by

$$\eta = \mathbf{A}^{[2]} \mathbf{b}^\omega - \mathbf{a}^{\omega\dagger} \mathbf{E}^{[3]} \mathbf{b}^\omega + \omega \mathbf{a}^{\omega\dagger} \mathbf{S}^{[3]} \mathbf{b}^\omega + \mathbf{B}^{[2]} \mathbf{a}^\omega \quad (43)$$

where the explicit expressions for the (super)matrices $\mathbf{A}^{[2]}$, $\mathbf{E}^{[3]}$, $\mathbf{S}^{[3]}$, and $\mathbf{B}^{[2]}$ may be found by differentiating with respect to the variational parameters X_i the lower order matrices ($\mathbf{A}^{[1]}$, $\mathbf{E}^{[2]}$, ...), using the rule

$$\frac{\partial \mathbf{D}(\mathbf{X})}{\partial X_j} = -[\mathbf{O}_j, \mathbf{D}]_S \quad (44)$$

which stems from the exponential parametrization of the density in eq 20.⁸⁶ These matrices are also collected in Table 1.

In terms of the AO matrices that constitute the building blocks of our computational procedure, the right-hand-side matrix (in the AO space) becomes

$$\eta = -[[\mathbf{S}, \mathbf{A}]_b, \mathbf{S}]_D + \mathbf{KDS} - \mathbf{SDK} + \omega \mathbf{S}[\mathbf{D}, [\mathbf{a}, \mathbf{b}^\dagger]_s]_s + [[\mathbf{S}, \mathbf{B}^\dagger]_a, \mathbf{S}]_D \quad (45)$$

with

$$\mathbf{K} = [\mathbf{S}, [\mathbf{F}, \mathbf{S}]_a]_b + [\mathbf{S}, \mathbf{G}([\mathbf{a}, \mathbf{D}])_s]_b + [\mathbf{S}, \mathbf{G}([\mathbf{b}^\dagger, \mathbf{D}]_s)]_a + \mathbf{G}([\mathbf{b}^\dagger, \mathbf{D}]_s, \mathbf{a}]_s) \quad (46)$$

where, for ease of notation, $\mathbf{a} \equiv \mathbf{a}^\omega$ and $\mathbf{b} \equiv \mathbf{b}^\omega$. Note that because of the structure of eq 42, also the response equation determining the orbital Lagrange multipliers can be solved using the iterative response solver of ref 74 with a modified right-hand-side.

Once the response vectors have been found the computation expression for the Verdet constant is obtained from the magnetic field derivative of the Lagrangian function

$$\frac{d\mathcal{L}^V}{dB_\gamma} = \mathbf{A}^{[1],\gamma\dagger} \mathbf{b}^\omega - \bar{\lambda}^\dagger [(\mathbf{E}^{[2],\gamma} - \omega \mathbf{S}^{[2],\gamma}) \mathbf{b}^\omega - \mathbf{B}^{[1],\gamma}] - \bar{\mathbf{X}}^\dagger \mathbf{E}^{[1],\gamma} \quad (47)$$

The $(2n + 1)$ and $(2n + 2)$ rules allow us to neglect the dependence on the magnetic field of the parameters (e.g., response vectors and Lagrange multipliers), whereas all other matrices in the Lagrangian depend on the magnetic field through the perturbation dependent basis set (i.e., the LAOs), and the explicit dependence on the magnetic field interaction in the perturbed Hamiltonian entering the Fock/Kohn–Sham matrix \mathbf{F} (hence $\mathbf{E}^{[2]}$ (see previous sections)). The explicit expressions for magnetic differentiated quantities are listed in Table 2.

Introducing the “perturbed” density matrices

$$\mathbf{D}^b = [\mathbf{b}, \mathbf{D}]_s \quad (48)$$

$$\mathbf{D}^a = [\mathbf{D}, \mathbf{a}^\dagger]_s \quad (49)$$

$$\mathbf{D}^{ab} = [[\mathbf{b}, \mathbf{D}]_s, \mathbf{a}^\dagger]_s + [\mathbf{D}, \bar{\mathbf{X}}^\dagger]_s \quad (50)$$

the final expression for the derivative of the linear response function with respect to the magnetic field in terms of the fundamental building blocks can be written

$$\begin{aligned} \frac{d\mathcal{L}^V}{dB_\gamma} \equiv \frac{d\alpha_{\alpha\beta}}{dB_\gamma} = & \text{Tr}\{\mathbf{D}^{ab} \mathbf{h}^\gamma + \mathbf{D}^{ab} \mathbf{G}^\gamma(\mathbf{D}) + \mathbf{G}^\gamma(\mathbf{D}^b) \mathbf{D}^a - \\ & \mathbf{b}[\mathbf{S}, \mathbf{A}^\gamma]_D - \mathbf{b}[\mathbf{S}, \mathbf{A}]_{D^\gamma} - \mathbf{b}[\mathbf{S}^\gamma, \mathbf{A}]_D - \mathbf{a}^\dagger[\mathbf{S}, \mathbf{B}^\gamma]_D - \\ & \mathbf{a}^\dagger[\mathbf{S}^\gamma, \mathbf{B}]_D - \mathbf{a}^\dagger[\mathbf{S}, \mathbf{B}]_{D^\gamma} - [\mathbf{D}, \bar{\mathbf{X}}^\dagger]_s \mathbf{F} - [[\mathbf{b}, \mathbf{D}]_s, \mathbf{a}^\dagger]_s \mathbf{F} - \\ & [\mathbf{D}^b, \mathbf{a}^\dagger]_s \mathbf{F} - [\mathbf{D}, \mathbf{a}^\dagger]_s \mathbf{G}(\mathbf{D}^b) - [\mathbf{b}, \mathbf{D}]_s \mathbf{G}(\mathbf{D}^a) - \\ & \omega(\mathbf{a}^\dagger \mathbf{S}[\mathbf{b}, \mathbf{D}]_s \mathbf{S} + \mathbf{a}^\dagger \mathbf{S}^\gamma \mathbf{D}^b \mathbf{S} + \mathbf{a}^\dagger \mathbf{S} \mathbf{D}^b \mathbf{S}^\gamma) - \mathbf{D}^\gamma \mathbf{G}(\mathbf{D}^{ab}) - \\ & [\mathbf{D}^\gamma, \bar{\mathbf{X}}^\dagger]_s \mathbf{F} - [[\mathbf{b}, \mathbf{D}^\gamma]_s, \mathbf{a}^\dagger]_s \mathbf{F} - [\mathbf{D}^\gamma, \mathbf{a}^\dagger]_s \mathbf{G}(\mathbf{D}^b) - \\ & [\mathbf{b}, \mathbf{D}^\gamma]_s \mathbf{G}(\mathbf{D}^a) - \omega \mathbf{a}^\dagger \mathbf{S}[\mathbf{b}, \mathbf{D}^\gamma]_s \} \quad (51) \end{aligned}$$

where $\mathbf{D}^\gamma = -\mathbf{DS}^\gamma \mathbf{D}$,⁸⁷ and \mathbf{S}^γ is the magnetic field derivative (in the γ direction) of the overlap matrix.

2.4.2. Standard Quadratic-Response Formulation of the Verdet Constant. If the LAO basis is not used eq 51 reduces to the simple expression

$$\begin{aligned} \frac{d\mathcal{L}^V}{dB_\gamma} = & \text{Tr}\{\mathbf{D}^{ab} \mathbf{h}^\gamma\} \\ = & \text{Tr}\{[[\mathbf{b}, \mathbf{D}]_s, \mathbf{a}^\dagger]_s \mathbf{h}^\gamma\} + \text{Tr}\{[\mathbf{D}, \bar{\mathbf{X}}^\dagger]_s \mathbf{h}^\gamma\} \quad (52) \end{aligned}$$

Table 2. Explicit Expressions for Magnetic Differentiated Quantities

$\mathbf{E}_{[1],\gamma}$	$= [\mathbf{S}^\gamma, \mathbf{F}]_{\text{D}} + [\mathbf{S}, \mathbf{F}^\gamma]_{\text{D}} + [\mathbf{S}, \mathbf{F}]_{\text{D}^\gamma}$
$\mathbf{E}_{[2],\gamma\mathbf{b}}$	$= [\mathbf{F}^\gamma, \mathbf{S}]_{[\mathbf{b}^\text{T}, \text{D}]_{\text{S}}} + [\mathbf{F}, \mathbf{S}^\gamma]_{[\mathbf{b}^\text{T}, \text{D}]_{\text{S}}} + [\mathbf{F}, \mathbf{S}]_{[\mathbf{b}^\text{T}, \text{D}^\gamma]_{\text{S}}} + [\mathbf{F}, \mathbf{S}]_{[\mathbf{b}^\text{T}, \text{D}]_{\text{S}^\gamma}} + [\mathbf{G}^{\gamma\text{T}}([\mathbf{D}, \mathbf{b}]_{\text{S}}), \mathbf{S}]_{\text{D}} + [\mathbf{G}^\text{T}([\mathbf{D}^\gamma, \mathbf{b}]_{\text{S}}), \mathbf{S}]_{\text{D}} + [\mathbf{G}^\text{T}([\mathbf{D}, \mathbf{b}]_{\text{S}^\gamma}), \mathbf{S}]_{\text{D}} + [\mathbf{G}^\text{T}([\mathbf{D}, \mathbf{b}]_{\text{S}}), \mathbf{S}]_{\text{D}^\gamma}$
$\mathbf{B}_{[1],\gamma}$	$= [\mathbf{S}^\gamma, \mathbf{B}^\text{T}]_{\text{D}} + [\mathbf{S}, (\mathbf{B}^\text{T})^\gamma]_{\text{D}} + [\mathbf{S}, \mathbf{B}^\text{T}]_{\text{D}^\gamma}$
$\mathbf{A}_{[1],\gamma}$	$= [(\mathbf{A}^\gamma)^\text{T}, \mathbf{S}]_{\text{D}} + [\mathbf{A}^\text{T}, \mathbf{S}^\gamma]_{\text{D}} + [\mathbf{A}^\text{T}, \mathbf{S}]_{\text{D}^\gamma}$
$\mathbf{S}_{[2],\gamma\mathbf{b}}$	$= \mathbf{S}^\gamma[\mathbf{D}, \mathbf{b}^\text{T}]_{\text{S}}\mathbf{S} + \mathbf{S}[\mathbf{D}^\gamma, \mathbf{b}^\text{T}]_{\text{S}}\mathbf{S} + \mathbf{S}[\mathbf{D}, \mathbf{b}^\text{T}]_{\text{S}^\gamma}\mathbf{S} + \mathbf{S}[\mathbf{D}, \mathbf{b}^\text{T}]_{\text{S}}\mathbf{S}^\gamma$

where \mathbf{h}^γ is now to the magnetic moment integral matrix on conventional atomic orbitals. Introducing the notation $\hat{\mathbf{C}} = \hat{h}^\gamma$ and the $\mathbf{C}^{[n]}$ matrices similarly to the $\mathbf{B}^{[n]}$,⁷⁵ that is, $\mathbf{C}^{[1]} = \mathbf{h}^\gamma \mathbf{D} \mathbf{S} - \mathbf{S} \mathbf{D} \mathbf{h}^\gamma$, we can rewrite eq 52 as (in supermatrix notation)

$$\frac{d\mathcal{L}^\gamma}{d\mathbf{B}_\gamma} = \mathbf{a}^{\omega^\dagger} \mathbf{C}^{[2]} \mathbf{b}^\omega + \bar{\mathbf{X}}^\dagger \mathbf{C}^{[1]} \quad (53)$$

and the second term of it according to

$$\begin{aligned} \bar{\mathbf{X}}^\dagger \mathbf{C}^{[1]} &= (\mathbf{A}^{[2]} \mathbf{b}^\omega - \mathbf{a}^{\omega^\dagger} \mathbf{E}^{[3]} \mathbf{b}^\omega + \omega \mathbf{a}^{\omega^\dagger} \mathbf{S}^{[3]} \mathbf{b}^\omega + \\ &\quad \mathbf{B}^{[2]} \mathbf{a}^{\omega^\dagger} \mathbf{E}^{[2]^{-1}} \mathbf{C}^{[1]}) \\ &= (\mathbf{b}^{\omega^\dagger} \mathbf{A}^{[2]} - \mathbf{b}^{\omega^\dagger} \mathbf{E}^{[3]} \mathbf{a}^\omega + \omega \mathbf{b}^{\omega^\dagger} \mathbf{S}^{[3]} \mathbf{a}^\omega + \\ &\quad \mathbf{a}^{\omega^\dagger} \mathbf{B}^{[2]}) \mathbf{E}^{[2]^{-1}} \mathbf{C}^{[1]} \end{aligned} \quad (54)$$

Introducing the response vector \mathbf{c}^0 for the static magnetic field

$$\mathbf{E}^{[2]} \mathbf{c}^0 = \mathbf{C}^{[1]} \quad (55)$$

we obtain the standard expression for the quadratic response function⁷⁵

$$\begin{aligned} \frac{d\mathcal{L}^\gamma}{d\mathbf{B}_\gamma} &= \mathbf{a}^{\omega^\dagger} \mathbf{C}^{[2]} \mathbf{b}^\omega + \mathbf{b}^{\omega^\dagger} \mathbf{A}^{[2]} \mathbf{c}^0 - \mathbf{b}^{\omega^\dagger} \mathbf{E}^{[3]} \mathbf{a}^\omega \mathbf{c}^0 + \\ &\quad \omega \mathbf{b}^{\omega^\dagger} \mathbf{S}^{[3]} \mathbf{a}^\omega \mathbf{c}^0 + \mathbf{a}^{\omega^\dagger} \mathbf{B}^{[2]} \mathbf{c}^0 \\ &= \langle \langle \hat{\mathbf{A}}; \hat{\mathbf{B}}, \hat{\mathbf{C}} \rangle \rangle_{\omega, 0} \end{aligned} \quad (56)$$

Equation 51 thus reduces to the standard expression obtained from conventional quadratic response. Formally, both eqs 51 and 52 follow a $n + 1$ rule (with respect to the electric field), inferior to the $2n + 1$ rule, since the Lagrangian multiplier $\bar{\mathbf{X}}$ is a second order quantity, depending simultaneously on both the α and β directions of the perturbing electric fields. Equation 56, as well as the method presented by Krykunov et al.,^{18,19} follow the superior $2n + 1$ rule. It should be noted, however, that the Verdet constant only requires crossed terms $\alpha \neq \beta \neq \gamma$, that is, one $\bar{\mathbf{X}}$ vector for each direction of the perturbing magnetic field, yielding to the same number of equations to be solved as in the $2n + 1$ case. Moreover, we do not need to solve any response equation that depends on the magnetic field.

2.4.3. Lagrangian Formulation of the \mathcal{B} Term. For the \mathcal{B} term, the Lagrangian functional is built as

$$\mathcal{L}^{\mathcal{B}} = \mathbf{A}^{[1]\dagger} \mathbf{b}^f - \bar{\lambda}^\dagger (\mathbf{E}^{[2]} - \omega_f \mathbf{S}^{[2]}) \mathbf{b}^f - \bar{\omega} (\mathbf{b}^{f\dagger} \mathbf{S}^{[2]} \mathbf{b}^f - 1) - \bar{\mathbf{X}}^\dagger \mathbf{E}^{[1]} \quad (57)$$

that is, from the expression for the transition moment, plus three constraint equations, namely the generalized eigenvalue

equation for the excited state vector \mathbf{b}^f (eq 26), its orthonormality condition (eq 27) and the optimization condition for the variational parameters \mathbf{X} (eq 39).

Imposing variationality with respect to the excitation vectors yields the equation for the ‘‘property’’ multipliers $\bar{\lambda}^\dagger$ and $\bar{\omega}$

$$\frac{\partial \mathcal{L}^{\mathcal{B}}}{\partial \mathbf{b}_j^f} = 0 \Leftrightarrow A_j^{[1]} = \bar{\lambda}_m (E_{mj}^{[2]} - \omega_f S_{mj}^{[2]}) + 2\bar{\omega} b_m^f S_{mj}^{[2]} \quad (58)$$

Equation 58 may be written in its adjoint form

$$(\mathbf{E}^{[2]} - \omega_f \mathbf{S}^{[2]}) \bar{\lambda} + 2\bar{\omega} \mathbf{S}^{[2]} \mathbf{b}^f = \mathbf{A}^{[1]} \quad (59)$$

and projected against $\mathbf{b}^{f\dagger}$ and its orthogonal complement space

$$\mathbf{P}^{f\dagger} = (\mathbf{I} - \mathbf{S}^{[2]} \mathbf{b}^f \mathbf{b}^{f\dagger}) \quad (60)$$

Using both $\mathbf{b}^{f\dagger} (\mathbf{E}^{[2]} - \omega_f \mathbf{S}^{[2]}) = 0$ and the orthonormality condition eq 27, these projections give, respectively

$$\bar{\omega} = \frac{1}{2} b_j^f A_j^{[1]} = \frac{1}{2} \mathbf{b}^{f\dagger} \mathbf{A}^{[1]} \quad (61)$$

and

$$\mathbf{P}^{f\dagger} ((\mathbf{E}^{[2]} - \omega_f \mathbf{S}^{[2]}) \bar{\lambda}) = \mathbf{P}^{f\dagger} \mathbf{A}^{[1]} \quad (62)$$

Thus, the $\mathbf{b}^{f\dagger}$ -projected equation, eq 61, uniquely determines the $\bar{\omega}$ multiplier, whereas the $\mathbf{P}^{f\dagger}$ -projected equation, eq 62, determines the multipliers $\bar{\lambda}$. Since $\mathbf{P}^{f\dagger} (\mathbf{E}^{[2]} - \omega_f \mathbf{S}^{[2]}) = (\mathbf{E}^{[2]} - \omega_f \mathbf{S}^{[2]}) \mathbf{P}^f$, where

$$\mathbf{P}^f = (\mathbf{I} - \mathbf{b}^f \mathbf{b}^{f\dagger} \mathbf{S}^{[2]}) \quad (63)$$

Equation 62 may be solved as

$$\mathbf{P}^{f\dagger} (\mathbf{E}^{[2]} - \omega_f \mathbf{S}^{[2]}) \mathbf{P}^f \bar{\lambda} = \mathbf{P}^{f\dagger} \mathbf{A}^{[1]} \quad (64)$$

Since we solve the above equation using an iterative procedure where trial vectors are added in pairs,⁷⁴ some modifications have to be introduced to solve these projected equations. These modifications are described in Appendix A, and the modified preconditioned conjugated gradient method displays the same convergence properties as the standard iterative procedure presented in ref 74. Note that the authors of ref 42 face a similar problem, with a singular equation, which is solved by a more straightforward projection, since the paired structure is not exploited. Note also that the corresponding equation in ref 42 (eq 24) is solved in a form with half the dimension of eq 64. However, this is done at the expense that squared excitations are obtained by solving sets of linear equations which have the squared conditioning number of eq 64 strongly impeding the convergence of the equations.

The variational condition with respect to the (nonredundant) orbital parameters X_j results in the response equation for the orbital Lagrangian multipliers $\bar{\mathbf{X}}$, which is clearly still

in the form of eq 42, but with a slightly different expression for the right-hand-side vector

$$\begin{aligned} \eta &= \mathbf{b}^{\dagger} \mathbf{A}^{[21]} - \mathbf{a}^{\omega \dagger} \mathbf{E}^{[31]} \mathbf{b}^f + \omega \mathbf{a}^{\omega \dagger} \mathbf{S}^{[31]} \mathbf{b}^f - \\ &\quad \bar{\omega} \mathbf{b}^{\dagger} \mathbf{S}^{[31]} \mathbf{b}^f + \frac{\partial \bar{\omega}}{\partial \mathbf{X}} (\mathbf{b}^{\dagger} \mathbf{S}^{[21]} \mathbf{b}^f - 1) \\ &= \mathbf{b}^{\dagger} \mathbf{A}^{[21]} + \omega \mathbf{a}^{\omega \dagger} \mathbf{S}^{[31]} \mathbf{b}^f - \mathbf{a}^{\omega \dagger} \mathbf{E}^{[31]} \mathbf{b}^f \end{aligned} \quad (65)$$

Equation 65 may be expressed in terms of the AO building blocks

$$\eta = -[[\mathbf{S}, \mathbf{A}]_{\mathbf{b}}, \mathbf{S}]_{\mathbf{D}} + \omega_f \mathbf{S}[\mathbf{D}, [\mathbf{a}, \mathbf{b}^{\dagger}]_{\mathbf{S}}]_{\mathbf{S}} + \mathbf{KDS} - \mathbf{SDK} \quad (66)$$

where $\mathbf{b} = \mathbf{b}^f$, and $\mathbf{a} = P^f(\lambda)$, that is, the AO matrix form of the projected solution of linear response equation in eq 62. The matrix \mathbf{K} is defined analogous to the one for the Verdet constant in eq 46, but with \mathbf{a} and \mathbf{b} given as above.

To obtain the \mathcal{B} term we finally consider the magnetic field derivative of the Lagrangian function

$$\left. \frac{d\mathcal{L}^{\mathcal{B}}}{dB_{\gamma}} \right|_{B_{\gamma}=0} = \mathbf{A}^{[1],\gamma \dagger} \mathbf{b}^f - \bar{\lambda}^{\dagger} (\mathbf{E}^{[2],\gamma} - \omega_f \mathbf{S}^{[2],\gamma}) \mathbf{b}^f + \bar{\omega} \mathbf{b}^{\dagger} \mathbf{S}^{[2],\gamma} \mathbf{b}^f - \bar{\mathbf{X}}^{\dagger} \mathbf{E}^{[1],\gamma} \quad (67)$$

The explicit expressions for magnetic differentiated quantities are listed in Table 2. Introducing, similar to what was done for the Verdet constant, the ‘‘perturbed’’ transition-density matrices

$$\mathbf{D}^f = [\mathbf{b}^f, \mathbf{D}]_{\mathbf{S}} \quad (68)$$

$$\mathbf{D}^a = [\mathbf{D}, \mathbf{a}]_{\mathbf{S}} \quad (69)$$

$$\mathbf{D}^{af} = [[\mathbf{b}^f, \mathbf{D}]_{\mathbf{S}}, \mathbf{a}^{\dagger}]_{\mathbf{S}} + [\mathbf{D}, \bar{\mathbf{X}}^{\dagger}]_{\mathbf{S}} \quad (70)$$

(where \mathbf{a} is the projected linear response vector) we write the final expression for the derivative of the dipole transition strength with respect to the magnetic field in terms of the fundamental building blocks as

$$\begin{aligned} \frac{d\mathcal{L}^{\mathcal{B}}}{dB_{\gamma}} &= \text{Tr}\{\mathbf{D}^{af} \mathbf{h}^{\gamma} + \mathbf{D}^{af} \mathbf{G}^{\gamma}(\mathbf{D}) + \mathbf{G}^{\gamma}(\mathbf{D}^f) \mathbf{D}^a - \\ &\quad \mathbf{b}^{\dagger} [\mathbf{S}, \mathbf{A}^{\gamma}]_{\mathbf{D}} - \mathbf{b}^{\dagger} [\mathbf{S}, \mathbf{A}]_{\mathbf{D}_{\gamma}} - \mathbf{b}^{\dagger} [\mathbf{S}^{\gamma}, \mathbf{A}]_{\mathbf{D}} - [\mathbf{D}, \bar{\mathbf{X}}^{\dagger}]_{\mathbf{S}_{\gamma}} \mathbf{F} - \\ &\quad [[\mathbf{b}^f, \mathbf{D}]_{\mathbf{S}_{\gamma}}, \mathbf{a}^{\dagger}]_{\mathbf{S}} \mathbf{F} - [\mathbf{D}^f, \mathbf{a}^{\dagger}]_{\mathbf{S}_{\gamma}} \mathbf{F} - [\mathbf{D}, \mathbf{a}^{\dagger}]_{\mathbf{S}_{\gamma}} \mathbf{G}(\mathbf{D}^f) - \\ &\quad [\mathbf{b}^f, \mathbf{D}]_{\mathbf{S}_{\gamma}} \mathbf{G}(\mathbf{D}^a) - \omega_f (\mathbf{a}^{\dagger} \mathbf{S}[\mathbf{b}^f, \mathbf{D}]_{\mathbf{S}_{\gamma}} \mathbf{S} + \mathbf{a}^{\dagger} \mathbf{S}^{\gamma} \mathbf{D}^f \mathbf{S} + \\ &\quad \mathbf{a}^{\dagger} \mathbf{S} \mathbf{D}^f \mathbf{S}^{\gamma}) - \mathbf{D}^{\gamma} \mathbf{G}(\mathbf{D}^{af}) - [\mathbf{D}^{\gamma}, \bar{\mathbf{X}}^{\dagger}]_{\mathbf{S}} \mathbf{F} - [[\mathbf{b}^f, \mathbf{D}^{\gamma}]_{\mathbf{S}}, \mathbf{a}^{\dagger}]_{\mathbf{S}} \mathbf{F} - \\ &\quad [\mathbf{D}^{\gamma}, \mathbf{a}^{\dagger}]_{\mathbf{S}} \mathbf{G}(\mathbf{D}^f) - [\mathbf{b}^f, \mathbf{D}^{\gamma}]_{\mathbf{S}} \mathbf{G}(\mathbf{D}^a) - \omega_f \mathbf{a}^{\dagger} \mathbf{S}[\mathbf{b}^f, \mathbf{D}^{\gamma}]_{\mathbf{S}} + \\ &\quad \bar{\omega} (\mathbf{b}^{\dagger} \mathbf{S}[\mathbf{b}^f, \mathbf{D}]_{\mathbf{S}_{\gamma}} \mathbf{S} + \mathbf{b}^{\dagger} \mathbf{S}^{\gamma} \mathbf{D}^f \mathbf{S} + \mathbf{b}^{\dagger} \mathbf{S} \mathbf{D}^f \mathbf{S}^{\gamma} + \mathbf{b}^{\dagger} \mathbf{S}[\mathbf{b}^f, \mathbf{D}^{\gamma}]_{\mathbf{S}}) \end{aligned} \quad (71)$$

where $\bar{\omega}$, which was defined in eq 61, and is computed as

$$\bar{\omega} = \frac{1}{2} \text{Tr}\{\mathbf{b}^{\dagger} [\mathbf{A}, \mathbf{S}]_{\mathbf{D}}\} \quad (72)$$

If a conventional atomic orbital basis is used instead on the LAO one, the above expression simply reduces to

$$\frac{d\mathcal{L}^{\mathcal{B}}}{dB_{\gamma}} = \text{Tr}\{\mathbf{D}^{af} \mathbf{h}^{\gamma}\} \quad (73)$$

which with methods similar to that used in section 2.4.2 can be shown to correspond to the standard expression of the Faraday \mathcal{B} term calculated as the first residue of the frequency dependent quadratic response functions.³⁴

3. Illustrative Results

The computational expressions for the two MOA properties have been implemented in a local, linear-scaling version of the DALTON code. The code has been used to compute Verdet constants of H₂, HF, CO, N₂O, and propane (CH₃CH₂CH₃) and the \mathcal{B} terms of pyrimidine, pyridine, and phosphabenzene. Selected augmented correlation-consistent basis sets^{88–90} were used. For the Verdet constant, calculations were performed at $\lambda = 400$ nm for all molecules except propane, for which $\lambda = 436$ has been used for a more straightforward comparison with experiment.^{91,92} At the DFT level the performance of 3 functionals was investigated, namely, the local density approximation LDA in the VWN5 parametrization,⁹³ B3LYP,⁹⁴ and CAMB3LYP.^{95,96} The results have been benchmarked against gauge-origin independent results obtained at the CCSD level using the methodology of refs 17 and 36. Experimentally derived equilibrium geometries were used for all systems except pyrimidine, pyridine, and phosphabenzene, for which we used the B3LYP/cc-pVTZ optimized structures of ref 36.

3.1. Verdet Constants. In Table 3, we report the results for the Verdet constant of H₂ obtained at the Hartree–Fock and DFT levels. H₂ is gauge-origin independent by symmetry, but it represents nonetheless a system of choice as we wish to monitor the effect of using LAOs in terms of improved convergence with respect to the basis set expansion. H₂ is also a good test system to benchmark the performance of the DFT potentials because we can compare with Full Configuration Interaction results, obtained by means of the CCSD/LAO implementation of ref 17.

Another system of choice for benchmark purposes is often HF. In this case, experimental results are not available; therefore, we limit the comparison of our Hartree–Fock and DFT values, given in Table 4, with previous gauge-origin independent CCSD results obtained in our group,¹⁷ as well as TD-DFT results reported by Ziegler and co-workers.

Our results for carbon monoxide are collected in Table 5 and compared with both experimental and TDDFT values from Krykunov et al.¹⁹ and with CCSD values obtained in this work.

The Verdet constants for both N₂O and propane have never been studied previously at any level of theory and are given in Tables 6 and 7, respectively. Experimental and CCSD results have also been given for N₂O, whereas for propane CCSD results have not been included, since they proved to be computationally too expensive to obtain.

We start from an analysis of the convergence of the computed results, which we illustrate in Figures 1–5. In general, one would expect the LAO basis improves the

Table 3. Magnetic Gradient of the Dipole Polarizability and Verdet Constant of H₂^a

method	basis	$\alpha_{yz}^{(m)}$	$\alpha_{yz}^{(m)}$		$V(\omega) \times 10^7$ a.u.		
			non-LAO	LAO	non-LAO	LAO	
Hartree–Fock	aug-cc-pVDZ	3.121966	4.721819	5.018943	0.40829	0.48696	
	aug-cc-pVTZ	3.67641	4.739413	4.861765	0.45592	0.464397	
	aug-cc-pVQZ	3.731024	4.781345	4.802397	0.460716	0.462175	
	aug-cc-pV5Z	3.736175	4.785383	4.788590	0.461174	0.461396	
	aug-cc-pV6Z	3.735484	4.785099	4.785857	0.461131	0.461183	
	d-aug-cc-pVDZ	3.609032	4.329553	5.464458	0.45236	0.47296	
	d-aug-cc-pVTZ	3.747092	4.801706	4.802905	0.46268	0.46277	
	d-aug-cc-pVQZ	3.738836	4.789761	4.788826	0.46157	0.46151	
	d-aug-cc-pV5Z	3.732110	4.783864	4.787649	0.46093	0.46119	
	d-aug-cc-pV6Z	3.730353	4.782889	4.787839	0.46080	0.46114	
	LDA	aug-cc-pVDZ	3.876146	5.502460	8.721509	0.5157	0.7389
		aug-cc-pVTZ	5.050625	6.448561	7.159842	0.6220	0.6713
		aug-cc-pVQZ	5.286503	6.640074	6.920677	0.64346	0.6629
		aug-cc-pV5Z	5.377983	6.710685	6.842713	0.65152	0.66067
aug-cc-pV6Z		5.420837	6.743145	6.795810	0.65526	0.65891	
daug-cc-pVDZ		5.278094	6.634525	7.403487	0.64278	0.69608	
daug-cc-pVTZ		5.449331	6.773410	6.794174	0.65834	0.65978	
daug-cc-pVQZ		5.449552	6.765238	6.769265	0.65778	0.65806	
daug-cc-pV5Z		5.448720	6.763791	6.762968	0.65766	0.65760	
daug-cc-pV6Z		5.449119	6.764166	6.764389	0.65770	0.65771	
B3LYP		aug-cc-pVDZ	3.383620	4.836236	7.424569	0.45248	0.63189
		aug-cc-pVTZ	4.338292	5.591631	6.063502	0.53792	0.57063
		aug-cc-pVQZ	4.511904	5.730192	5.890203	0.55355	0.56464
		aug-cc-pV5Z	4.570646	5.774623	5.838328	0.55866	0.56308
	aug-cc-pV6Z	4.592076	5.791338	5.813249	0.56056	0.56208	
	daug-cc-pVDZ	4.446419	5.665185	6.543317	0.54677	0.60764	
	daug-cc-pVTZ	4.603829	5.810628	5.826199	0.56231	0.56339	
	daug-cc-pVQZ	4.605892	5.803541	5.804439	0.56189	0.56195	
	daug-cc-pV5Z	4.603208	5.800324	5.799311	0.56157	0.56150	
	daug-cc-pV6Z	4.601501	5.798986	5.799814	0.56142	0.56148	

^a Frequency $\omega = 0.11391$ a.u. FCI/aug-cc-pV6Z: 0.45565×10^{-7} a.u. (non-LAO); 0.45578×10^{-7} a.u. (LAO)¹⁷ Experiment: 0.501×10^{-7} a.u.⁹²

Table 4. Magnetic Gradient of the Dipole Polarizability and Verdet Constant of Hydrogen Fluoride^a

method	basis	$\alpha_{xyz}^{(m)}$	$\alpha_{yz}^{(m)}$		$V(\omega) \times 10^7$ a.u.		
			non-LAO	LAO	non-LAO	LAO	
Hartree–Fock	aug-cc-pVDZ	1.5042	1.6310	2.1578	0.16518	0.20169	
	aug-cc-pVTZ	1.7011	1.8510	2.2167	0.18725	0.21260	
	aug-cc-pVQZ	1.8402	2.0249	2.2130	0.20413	0.21717	
	aug-cc-pV5Z	1.9054	2.0939	2.2025	0.21117	0.21870	
	d-aug-cc-pVDZ	1.9379	2.1908	2.1591	0.21901	0.21683	
	d-aug-cc-pVTZ	1.9782	2.1751	2.1644	0.21932	0.21857	
	d-aug-cc-pVQZ	1.9802	2.1753	2.1729	0.21940	0.21924	
	d-aug-cc-pV5Z	1.9794	2.1751	2.1740	0.21936	0.21929	
	LDA	aug-cc-pVDZ	3.98498	2.982455	4.185478	0.34482	0.42822
		aug-cc-pVTZ	4.116581	3.335878	4.329115	0.37389	0.44273
		aug-cc-pVQZ	4.444917	3.740532	4.448401	0.41331	0.46238
		aug-cc-pV5Z	4.64244	3.960753	4.451840	0.43542	0.46946
		d-aug-cc-pVDZ	4.89831	4.29881	4.46862	0.46772	0.47949
		d-aug-cc-pVTZ	5.01883	4.29655	4.312476	0.47174	0.47285
d-aug-cc-pVQZ		5.03779	4.345523	4.345219	0.47580	0.47578	
d-aug-cc-pV5Z		5.03777	4.353641	4.352530	0.47636	0.47628	
B3LYP		aug-cc-pVDZ	3.254867	2.660440	3.674269	0.29721	0.36748
		aug-cc-pVTZ	3.396116	2.948357	3.779644	0.32206	0.37968
		aug-cc-pVQZ	3.629149	3.261189	3.831202	0.35182	0.39133
		aug-cc-pV5Z	3.782167	3.435562	3.816149	0.36921	0.39559
		d-aug-cc-pVDZ	3.971842	3.706431	3.804644	0.39456	0.40136
		d-aug-cc-pVTZ	4.071619	3.701163	3.709247	0.39765	0.39821
	d-aug-cc-pVQZ	4.071138	3.727312	3.724679	0.39945	0.39902	
	d-aug-cc-pV5Z	4.072300	3.732808	3.729654	0.39986	0.39965	

^a Frequency $\omega = 0.11391$ a.u. Origin on H. CCSD/aug-cc-pVTZ: 0.26473×10^{-7} a.u. (non-LAO); 0.30883×10^{-7} a.u. (LAO)¹⁷ CCSD/d-aug-cc-pVTZ: 0.31936×10^{-7} a.u. (non-LAO); 0.32126×10^{-7} a.u. (LAO)¹⁷ LDA/ET3/LAO: 0.53030×10^{-7} a.u.; SAOP/ET3/LAO: 0.30134×10^{-7} a.u.¹⁹

convergence of the results because the orbitals are correct to first order in the applied magnetic field. Indeed, the differences, when increasing the cardinal number from $X = D$ to higher X , are significantly smaller when using the LAO basis for all methods/functionals considered, than when using conventional basis sets. The aug-cc-pVTZ/

LAO basis stands out as a reasonable compromise between accuracy and computational cost. Double augmentation clearly makes the usage of LAO somewhat unnecessary, especially for $X \geq 3$; on the other hand, one would seldom use a doubly augmented basis set for large systems, since they often results in linear dependences and numerical

Table 5. Magnetic Gradient of the Dipole Polarizability and Verdet Constant of Carbon Monoxide^a

method	basis	$\alpha_{xyz}^{(m)}$	$\alpha_{yzx}^{(m)}$		$V(\omega) \times 10^7$ a.u.		
			non-LAO	LAO	non-LAO	LAO	
Hartree–Fock	aug-cc-pVDZ	11.055792	3.616679	4.273038	0.63384	0.67934	
	aug-cc-pVTZ	11.406121	4.246722	4.424964	0.68965	0.70201	
	aug-cc-pVQZ	11.522226	4.436302	4.471400	0.70682	0.70925	
	aug-cc-pV5Z	11.546578	4.464267	4.486325	0.70960	0.71113	
	daug-cc-pVDZ	11.472916	4.441526	4.374207	0.70547	0.70081	
	daug-cc-pVTZ	11.529756	4.485288	4.474452	0.71047	0.70972	
	daug-cc-pVQZ	11.553393	4.490153	4.489820	0.71163	0.71161	
	daug-cc-pV5Z	11.556813	4.487931	4.486967	0.71159	0.71153	
	LDA	aug-cc-pVDZ	15.983420	6.836289	7.693072	1.02778	1.08716
		aug-cc-pVTZ	16.423412	7.438752	7.699148	1.08447	1.10283
aug-cc-pVQZ		16.603980	7.679665	7.754938	1.10774	1.11296	
aug-cc-pV5Z		16.644201	7.726884	7.770670	1.11240	1.11545	
B3LYP	aug-cc-pVDZ	13.936334	5.943637	6.815239	0.89496	0.95537	
	aug-cc-pVTZ	14.414739	6.587390	6.869864	0.95616	0.97574	
	aug-cc-pVQZ	14.583939	6.834959	6.915444	0.97918	0.98476	
	aug-cc-pV5Z	14.621654	6.886413	6.931134	0.98406	0.98717	
	daug-cc-pVDZ	14.501710	6.931655	6.867495	0.98304	0.97859	
	daug-cc-pVTZ	14.614166	6.934587	6.926226	0.98714	0.98656	
CAMB3LYP	daug-cc-pVQZ	14.644941	6.941044	6.942348	0.98865	0.98874	
	daug-cc-pV5Z	14.644213	6.940722	6.938852	0.98860	0.98847	
	aug-cc-pVDZ	13.553915	5.530387	6.423219	0.85306	0.91495	
	aug-cc-pVTZ	14.034989	6.194255	6.475651	0.91575	0.93525	
	aug-cc-pVQZ	14.206308	6.519579	6.448183	0.93929	0.94423	
	aug-cc-pV5Z	14.244126	6.497478	6.536160	0.94402	0.94670	

^a Frequency $\omega = 0.11391$ a.u. Gauge-origin on C atom. CCSD/aug-cc-pVTZ: 0.83141×10^{-7} a.u. (non-LAO); 0.85749×10^{-7} a.u. (LAO) CCSD/aug-cc-pVQZ: 0.83560×10^{-7} a.u. (non-LAO); 0.85122×10^{-7} a.u. (LAO) LDA/ET3(LAO): 1.083409×10^{-7} a.u.¹⁹ SAOP/ET3(LAO): 0.9156984×10^{-7} a.u.¹⁹ Experiment: 0.895×10^{-7} a.u.⁹²

Table 6. Magnetic Gradient of the Dipole Polarizability and Verdet Constant of Nitrous Oxide, N₂O^a

method	basis	$\alpha_{xyz}^{(m)}$	$\alpha_{yzx}^{(m)}$		$V(\omega) \times 10^7$ a.u.	
			non-LAO	LAO	non-LAO	LAO
Hartree–Fock	aug-cc-pVDZ	5.20186	3.34222	5.72895	0.41193	0.57737
	aug-cc-pVTZ	5.75131	4.61477	5.36470	0.51919	0.57117
	aug-cc-pVQZ	5.99701	5.01297	5.18182	0.55530	0.56701
	aug-cc-pV5Z	6.05247	5.08413	5.16840	0.56216	0.56800
LDA	aug-cc-pVDZ	7.46008	2.26474	5.75498	0.41552	0.65744
	aug-cc-pVTZ	8.02973	3.72087	5.01409	0.53619	0.62583
	aug-cc-pVQZ	8.37503	4.24664	4.68324	0.58460	0.61486
B3LYP	aug-cc-pV5Z	8.469322	4.391367	4.619600	0.59790	0.61372
	aug-cc-pVDZ	6.945838	2.406820	5.866543	0.40754	0.64735
	aug-cc-pVTZ	7.484019	3.831389	5.168035	0.52494	0.61758
	aug-cc-pVQZ	7.835584	4.382733	4.828869	0.57567	0.60625
CAMB3LYP	aug-cc-pV5Z	7.939933	4.537069	4.755828	0.58965	0.60481
	aug-cc-pVDZ	6.749242	2.439556	5.668062	0.402998	0.62678
	aug-cc-pVTZ	7.254741	3.871418	5.038626	0.519765	0.60067
	aug-cc-pVQZ	7.593739	4.407516	4.743719	0.568672	0.59198

^a Frequency 0.11391 a.u. Gauge-origin on the O atom. CCSD/aug-cc-pVQZ: 0.47258×10^{-7} a.u. (non-LAO); 0.54402×10^{-7} a.u. (LAO) Experiment: 0.596×10^{-7} a.u.⁹²

Table 7. Magnetic Gradient of the Dipole Polarizability and Verdet Constant of Propane^a

method	basis	$\epsilon_{\alpha\beta\gamma}\alpha_{\alpha\beta\gamma}^{(m)}$	$V(\omega) \times 10^7$ a.u.		
			non-LAO	LAO	
Hartree–Fock	aug-cc-pVDZ	107.493	114.951	1.7089	1.8274
	aug-cc-pVTZ	113.752	115.109	1.8083	1.8299
	aug-cc-pVQZ	114.873	115.300	1.8262	1.8330
	d-aug-cc-pVDZ	115.287	115.947	1.8328	1.8432
	d-aug-cc-pVTZ	115.494	115.506	1.8360	1.8362
	d-aug-cc-pVQZ	115.502	115.506	1.8360	1.8362
LDA	aug-cc-pVDZ	149.344	167.233	2.3742	2.6585
	aug-cc-pVTZ	162.149	166.731	2.5777	2.6506
	aug-cc-pVQZ	165.502	167.411	2.6310	2.6614
	d-aug-cc-pVDZ	168.806	170.105	2.6836	2.7042
B3LYP	d-aug-cc-pVTZ	168.255	168.259	2.6748	2.6749
	aug-cc-pVDZ	132.217	147.596	2.1019	2.3258
	aug-cc-pVTZ	143.360	146.686	2.2790	2.3319
	aug-cc-pVQZ	145.816	147.166	2.3181	2.3396

^a Wavelength 436 nm. Origin at the center of mass. Experiment: 2.34×10^{-7} a.u.⁹²

instabilities, in all such cases the LAO basis stands as an effective choice to improve the convergence.

The other advantage of using the LAO basis is the gauge-origin independence of the results for any choice of basis

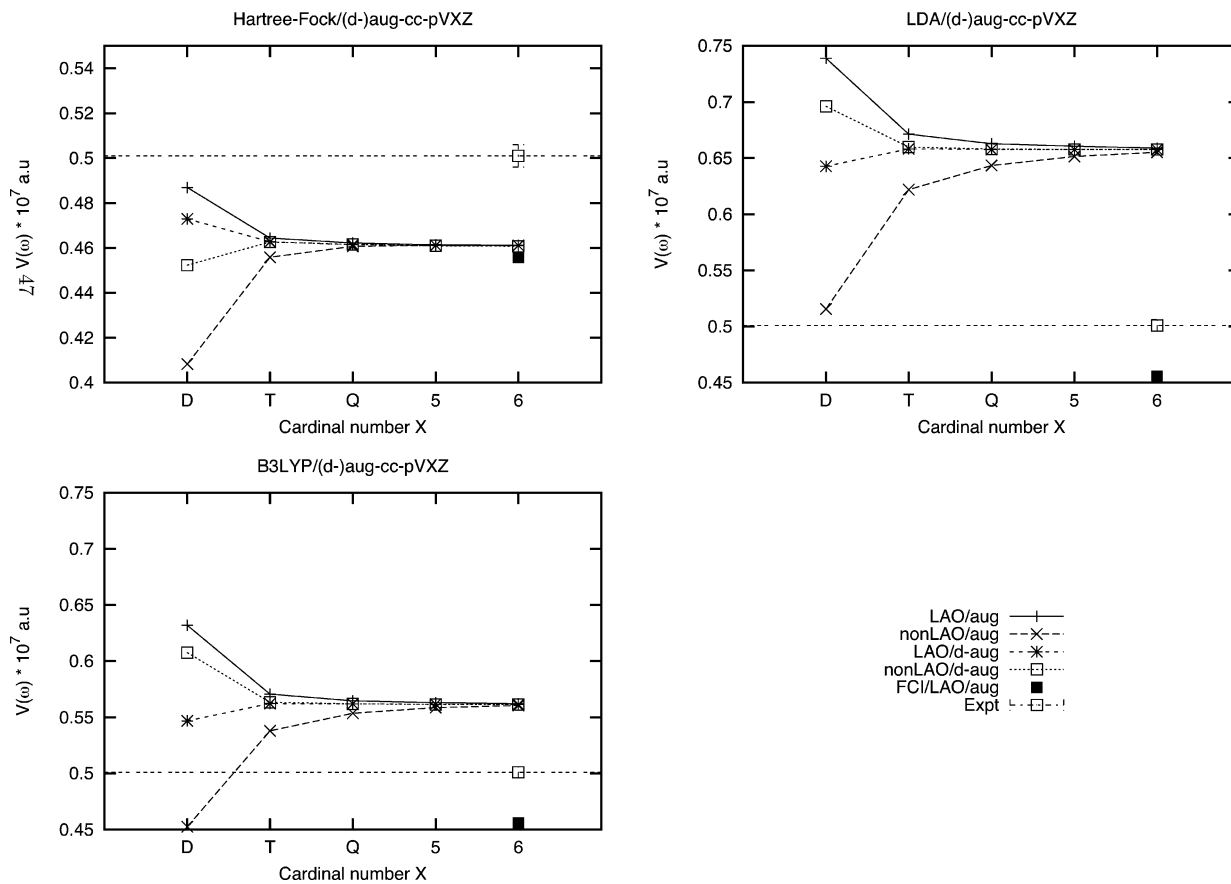


Figure 1. H₂: Convergence of the Verdet constant with the basis set.

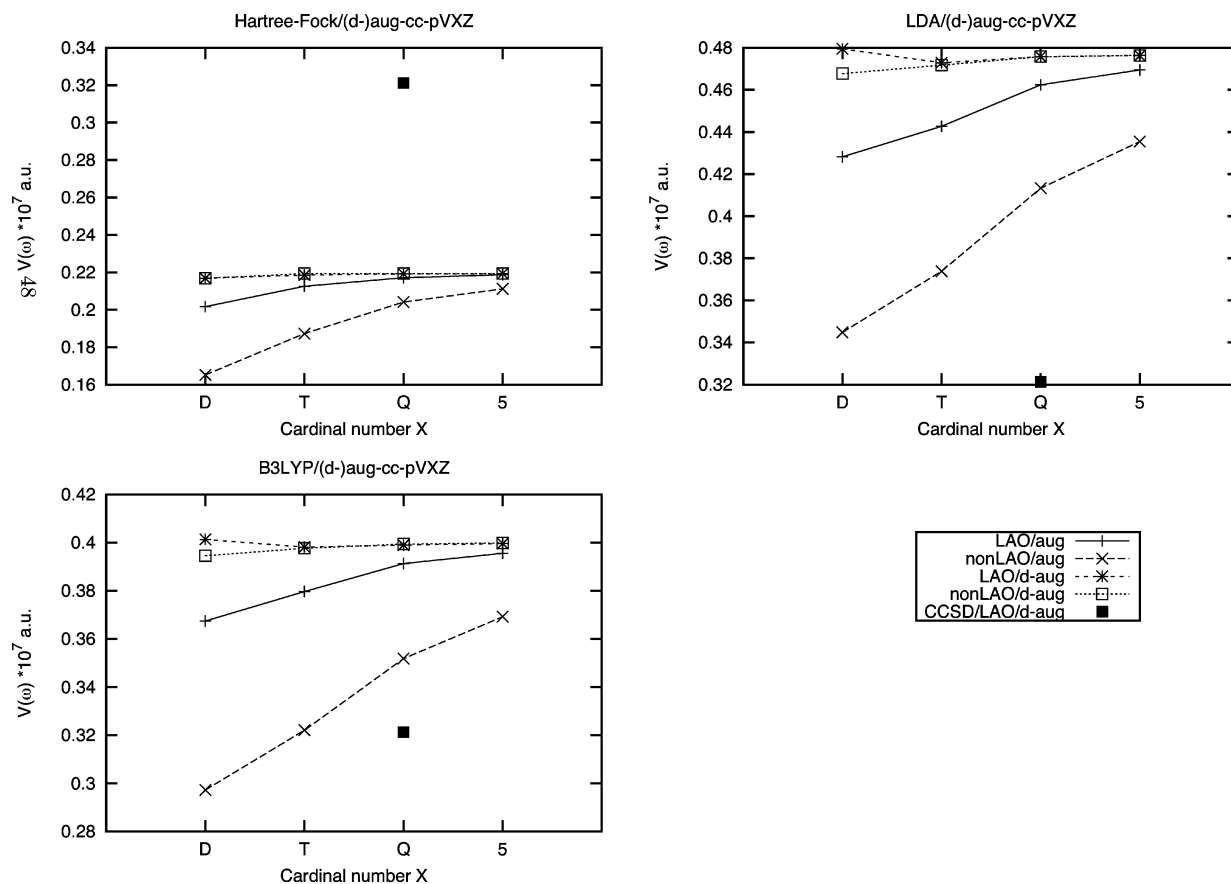


Figure 2. HF: Convergence of the Verdet constant with the basis set.

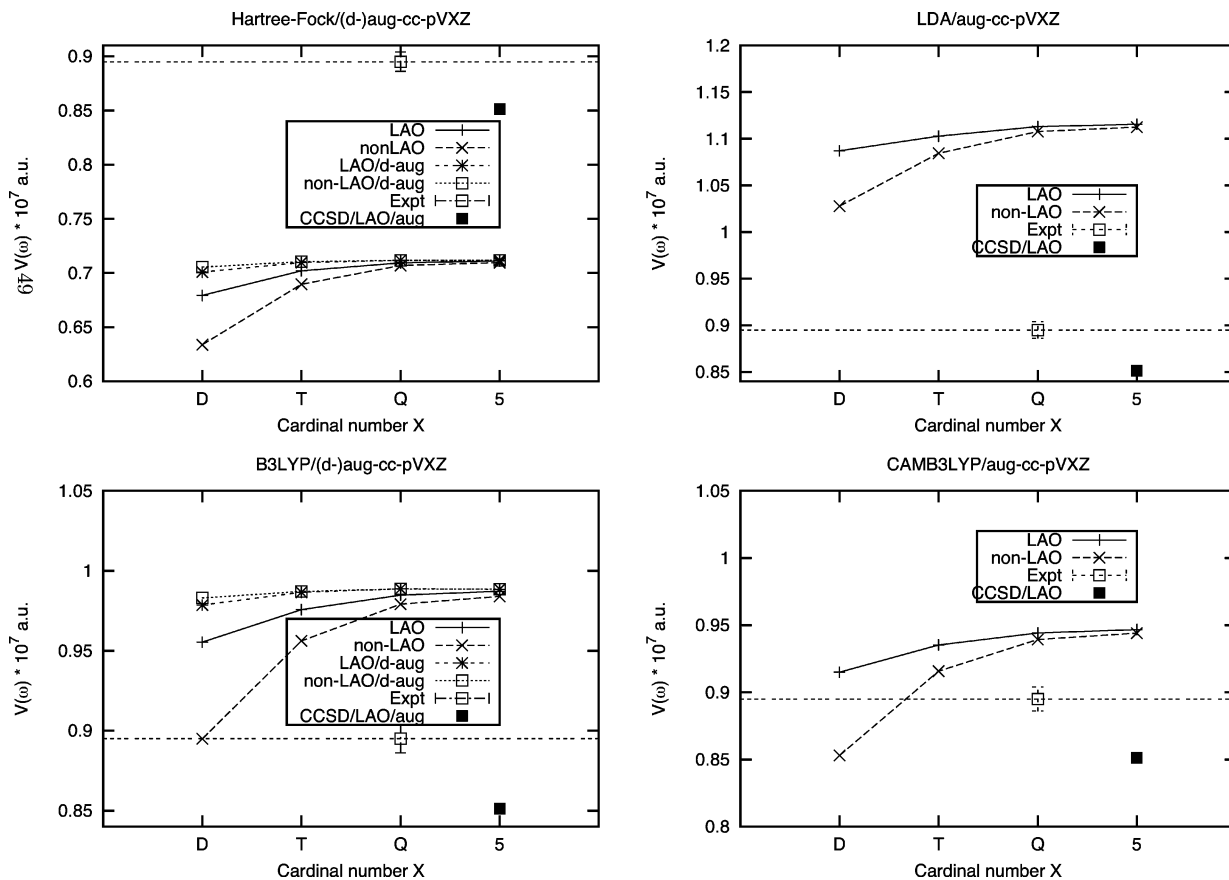


Figure 3. CO: Convergence of the Verdet constant with the basis set.

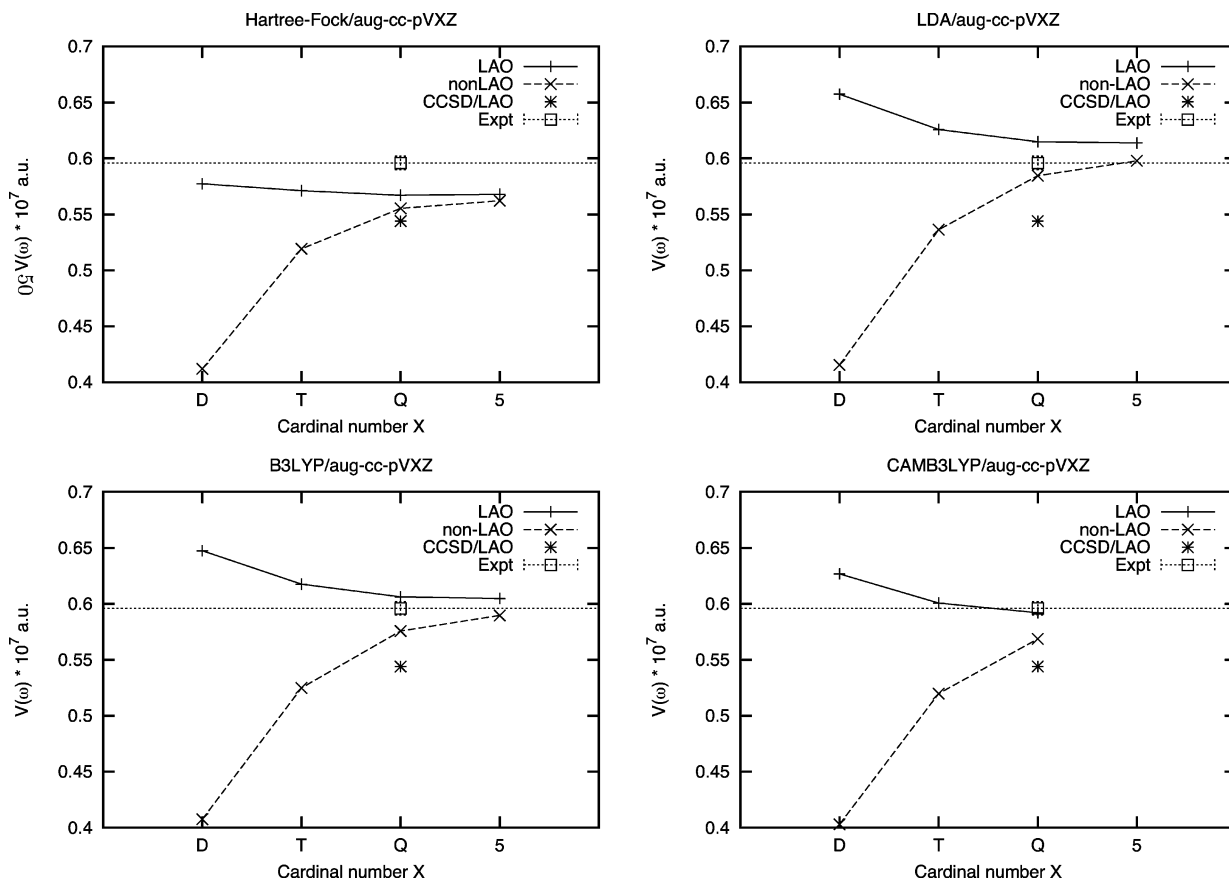


Figure 4. N₂O: Convergence of the Verdet constant with the basis set.

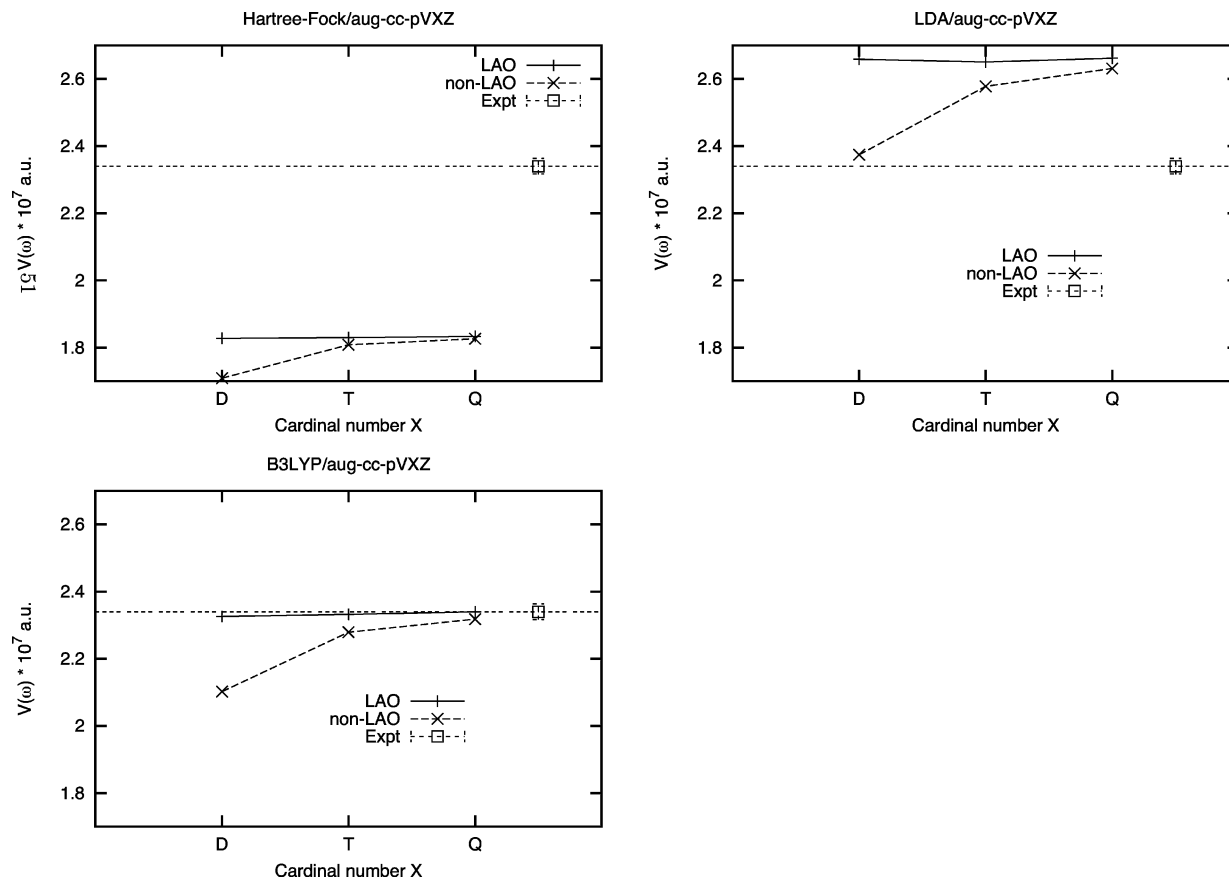


Figure 5. Propane: Convergence of the Verdet constant with the basis set.

set. The dependence of the results upon a shift of the origin of the coordinate system can, in the conventional basis set, be expressed as the difference between the value obtained at a given origin and the one obtained shifting the origin to a different point. Such difference can be written as product of the shift vector with an appropriate response function,⁴⁹ and it is expected to diminish when enlarging the basis set. We have estimated the gauge origin dependence of the results for CO, N₂O, and propane when using conventional (AO) basis sets by performing a few calculation at two different origins for each system. The overall gauge origin dependence effect was found to be rather modest with a reasonable choice of origin (below 1%) but it can be expected to become larger for extended systems.

Comparing with the experimental results at wavelengths according to refs 91 and 92, for which the authors claim an accuracy of 1%, and with our CCSD and previous DFT results from the literature, we notice that (i) For H₂ all DFT functionals here considered significantly overestimate the Verdet constant, both with respect to the benchmark FCI value, and the experimental value.^{91,92} The Hartree–Fock/LAO results are, as previously observed for conventional basis sets,¹⁵ rather close to the FCI results. The difference between experimental and FCI results can be attributed to the neglect of ro-vibrational averaging, which is of the order of 10%.^{97,98} (ii) For HF, the Hartree–Fock result is well below the CCSD value, whereas both LDA and B3LYP results are larger. Our LDA results are smaller than the LDA/ET3(LAO) ones by Krykunov and co-workers.¹⁹ (iii) The Verdet constant of CO is underestimated by the Hartree–Fock

method and overestimated by all DFT functionals, as also observed by Krykunov and co-workers.¹⁹ The CCSD results are, on the other hand, always smaller than the experimental value. None of the methods, including CCSD, yields results that fall within the 1% accuracy of the experimental value (notice that no ro-vibrational averaging has been performed but the vibrational contribution is of the order of 1%.)⁹⁸ This is also in line with previous findings, like, for instance, for N₂, C₂H₂ and CH₄.¹⁶ (iv) In N₂O all methods, apart from LDA/LAO tend to underestimate the Verdet constant. All methods also give Verdet constants that are higher than the CCSD values. The latter are, as in the previous cases, smaller than the experimental values. Also in this case all methods yield Verdet constants, which are outside the error bars of the experimental value (again neglecting ro-vibrational averaging), but to a much lesser extent than what was observed for CO. (v) In the case of propane, Hartree–Fock is once more yielding Verdet constants which are too low and outside the 1% experimental accuracy, whereas LDA results are too high. The B3LYP results, on the other hand, fall within the 1% experimental accuracy range, despite the neglect of vibrational contribution. Mort and Autschbach showed that larger molecules like benzene, toluene, *p*-xylene, and *o*-xylene also have a sizable vibrational contribution and the vibrational contribution must be taken into account for high accuracy calculations of the Verdet constant.

3.2. β Terms of Pyrimidine, Pyridine, and Phosphabenzene. The MCD spectra of pyrimidine, pyridine and phosphabenzene were the subject of a recent investigation at the CCSD/LAO level carried out in our group.³⁶ It was there

Table 8. Pyrimidine: Excitation Energies ω_f (eV), Oscillator Strengths f , and \mathcal{B} Terms ($10^{-3} \text{ D}^2\mu_B \text{ cm}$) for the First 5 Dipole-Allowed Electronic (Singlet) Transitions from the Ground State^a

method/basis	symmetry	ω_f	f	$\mathcal{B}(0 \rightarrow f)$	
				non-LAO	LAO
Hartree-Fock/aug-cc-pVDZ	1 B ₁	5.739	0.0072	-0.105	-0.102
	1 B ₂	6.175	0.0837	1.732	1.751
	1 A ₁	6.508	0.0234	-1.877	-1.895
	2 B ₁	7.364	0.0241	-0.042	-0.072
	2 A ₁	7.980	0.5536	5.79	5.91
LDA(VWN5)/aug-cc-pVDZ	1 B ₁	3.668	0.0034	-0.050	-0.049
	2 B ₁	5.155	0.0045	-0.0057	-0.0047
	1 B ₂	5.551	0.0002	0.0140	0.0132
	2 B ₂	5.599	0.0300	0.3087	0.3143
	1 A ₁	6.219	0.0076	-0.806	-1.866
B3LYP/aug-cc-pVDZ	1 B ₁	4.259	0.0046	-0.058	-0.057
	1 B ₂	5.732	0.0399	0.539	0.544
	2 B ₁	5.959	0.0049	-0.104	-0.101
	2 B ₂	6.154	0.0006	0.0138	0.0138
	1 A ₁	6.516	0.0318	-0.656	-0.657
B3LYP/aug-cc-pVTZ	1 B ₁	4.2584	0.0045	-0.057	-0.057
	1 B ₂	5.7240	0.0408	0.550	0.551
	2 B ₁	5.9563	0.0048	-0.100	-0.099
	2 B ₂	6.1562	0.0007	0.0151	0.0152
	1 A ₁	6.4981	0.0311	-0.667	-0.664
CAMB3LYP/aug-cc-pVDZ	1 B ₁	4.548	0.0054	-0.0670	-0.0654
	1 B ₂	5.799	0.0455	0.522	0.528
	2 B ₁	6.301	0.0057	-0.0717	-0.0688
	1 A ₁	6.651	0.0419	-0.866	-0.872
	2 B ₂	6.836	0.0044	0.0284	0.0275
CCSD/aug-cc-pVDZ-CM ³⁶	1 B ₁	4.64	0.006		-0.068
	1 B ₂	5.51	0.028		0.210
	2 B ₁	6.51	0.006		-0.055
	2 B ₂	6.68	0.008		0.017
	1 A ₁	6.98	0.027		-0.267
experiment ⁹⁹	1 B ₁	4.22	0.007		-0.06
	1 B ₂	5.21	0.03		0.2
experiment ⁶³	1 B ₁	4.29	0.0073		-0.076
	1 B ₂	5.17	0.033		0.24

^a Comparison of Hartree-Fock, LDA, B3LYP, CAMB3LYP and CCSD results. The molecule lies on the yz plane, with the C_2 symmetry axis along z .

noticed how for nonoverlapping bands (e.g., for pyrimidine) the \mathcal{B} terms calculated at the CCSD/LAO level are in good agreement with the corresponding experimental values, whereas for overlapping bands (as seen, e.g., for pyridine and phosphabenzene) significant cancellation effects between positive and negative contributions may occur, which results in a reduced magnitude of the experimentally derived \mathcal{B} terms as well as in a shift of the MCD peak maxima compared to the position of the peak maxima in the UV spectrum from which the excitation energies are usually derived. Note that the MCD spectra of pyridine and pyrimidine were also very recently investigated at the DFT level by Seth et al.⁴²

In Table 8, we collect the results for the excitation energies, oscillator strengths and \mathcal{B} terms of MCD relative to the first 5 dipole-allowed transitions of pyrimidine obtained at the Hartree-Fock and DFT level, together with the CCSD/LAO results from reference 36, and experimental results from references 63 and 69. Analogous results are collected in Table 9 for phosphabenzene and Table 10 for pyridine.

For all three molecules, a rather small difference between LAO and AO results for the \mathcal{B} terms is observed, at least for the transition considered here. Another general observation is the modest basis set effect for the low lying states (first 3–4 excitations), as it can be seen comparing the results for both excitation energies, oscillator strengths, and \mathcal{B} terms obtained using the aug-cc-pVTZ basis instead of the aug-cc-pVDZ in pyrimidine (see Table 8) and phosphabenzene (see Table 9). Such small differences hardly affect the

qualitative (and quantitative) features of the simulated spectra (see below). Because of the increased Rydberg character of the higher excited states, both the effect of an increased basis set and the effect of using London orbitals are more pronounced.

3.2.1. Pyrimidine. For pyrimidine, the first dipole-allowed excited state is assigned as B₁ by all methods, in accordance with the experimental assignment, but far too high in energy in the Hartree-Fock case and too low for LDA. The second state is assigned as B₂ by most methods, with the sole exception of LDA which yields a B₁ state. All methods yield negative \mathcal{B} terms for all states of B₁ symmetry, as well as the first state of A₁ symmetry, and positive \mathcal{B} terms for the states of B₂ symmetry, yet with magnitudes that vary significantly from method to method. Following the methodology adopted in ref 36, we show in Figure 6 the simulated MCD spectrum of pyrimidine obtained by associating to each vertical excitation energy and \mathcal{B} term, obtained using the CAMB3LYP (left panel) and the B3LYP (right panel) functionals, a Gaussian line shape function and convoluting over the entire frequency region. The CCSD curve is also plotted, together with a simulated experimental spectrum (labeled Expt.) constructed from the experimentally derived \mathcal{B} terms from ref 99. As it can be observed both DFT functionals display the correct qualitative features, but the absolute intensity differs significantly compared to both the CCSD values and experiment, with quite strong signals for the second

Table 9. Phosphabenzene: Excitation Energies ω_f (eV), Oscillator Strengths f , and \mathcal{B} Terms ($10^{-3} D^2\mu_B$ cm) for the First 6 Dipole-Allowed Electronic (Singlet) Transitions from the Ground State^a

method/basis	symmetry	ω_f	f	$\mathcal{B}(0 \rightarrow f)$	
				non-LAO	LAO
Hartree-Fock/aug-cc-pVDZ	1 A ₁	4.845	0.1048	1.017	1.029
	1 B ₂	5.105	0.0018	-0.979	-0.981
	1 B ₁	5.708	0.0214	0.0796	0.0782
	2 B ₁	6.274	0.0070	0.0674	0.0923
	2 B ₂	6.343	0.3149	3.588	3.615
	2 A ₁	6.810	0.6114	-2.512	-2.977
LDA(VWN5)/aug-cc-pVDZ	1 B ₁	4.520	0.0148	-0.237	-0.240
	1 B ₂	4.597	0.0010	0.292	0.296
	1 A ₁	5.061	0.1223	0.312	0.316
	2 B ₁	5.848	0.0011	-0.001	-0.007
	2 B ₂	5.860	0.2242	1.964	2.060
	2 A ₁	6.056	0.1223	-0.0014	-0.036
B3LYP/aug-cc-pVDZ	1 B ₂	4.728	0.0008	0.359	0.361
	1 B ₁	4.873	0.0173	-0.655	-0.656
	1 A ₁	5.051	0.1311	0.711	0.724
	2 B ₁	5.803	0.0012	0.0098	0.0128
	2 B ₂	5.932	0.2465	1.996	2.020
	2 A ₁	6.425	0.3042	-1.121	-1.332
B3LYP/aug-cc-pVTZ	1 B ₂	4.707	0.0012	0.475	0.474
	1 B ₁	4.867	0.0170	-0.771	-0.778
	1 A ₁	5.016	0.1291	0.676	0.685
	2 B ₁	5.776	0.0015	0.0159	0.0176
	2 B ₂	5.891	0.2363	2.026	2.022
	2 A ₁	6.392	0.3331	-1.676	-2.087
CAMB3LYP/aug-cc-pVDZ	1 B ₂	4.825	0.0018	0.7085	0.7122
	1 B ₁	5.058	0.0194	-3.342	-3.337
	1 A ₁	5.098	0.1419	2.970	2.973
	2 B ₂	6.052	0.2671	2.231	2.259
	2 B ₁	6.206	0.0031	0.0174	0.0259
	2 A ₁	6.581	0.4558	-1.390	-1.436
CCSD/aug-cc-pVDZ-CM ³⁶	1 B ₂	4.55	0.001	0.157	0.157
	1 B ₁	5.16	0.019	-0.438	-0.438
	1 A ₁	5.46	0.163	0.735	0.735
	2 B ₁	6.18	0.005	0.042	0.042
	2 B ₂	6.36	0.259	2.351	2.351
	2 A ₁	6.78	0.343	-4.795	-4.795
experiment ¹⁰⁰	1 B ₂	4.22		0.11	0.11
	1 B ₁	4.71		-0.10	-0.10
	1 A ₁	5.21	0.16		0.56

^a Comparison of Hartree-Fock, LDA, B3LYP, CAMB3LYP and CCSD results. The molecule lies on the yz plane, with the C_2 symmetry axis along z .

Table 10. Pyridine: Excitation Energies ω_f (eV), Oscillator Strengths f and \mathcal{B} Terms ($10^{-3} D^2\mu_B$ cm) for the First 6 Dipole-Allowed Electronic (Singlet) Transitions from the Ground State^a

method/basis	symmetry	ω_f	f	$\mathcal{B}(0 \rightarrow f)$	
				non-LAO	LAO
B3LYP/aug-cc-pVDZ	1 B ₁	4.765	0.0038	-0.0408	-0.0394
	1 B ₂	5.473	0.0402	0.3810	0.3858
	1 A ₁	6.162	0.0006	0.0215	0.0213
	2 A ₁	6.242	0.0167	-0.6201	-0.6275
	2 B ₂	6.654	0.0213	-0.7926	-0.8168
	2 B ₁	6.7988	0.0317	5.764	4.470
CAMB3LYP/aug-cc-pVDZ	3 A ₁	6.7997	0.0278	-5.179	-4.027
	1 B ₁	5.037	0.0044	-0.0474	-0.0458
	1 B ₂	5.542	0.0458	0.4037	0.4087
	1 A ₁	6.358	0.0183	-0.7004	-0.7082
	2 A ₁	6.869	0.0102	-0.0810	-0.0838
	2 B ₁	7.226	0.0453	-0.0109	-0.6836
CCSD/aug-cc-pVDZ-CM ³⁶	2 B ₂	7.295	0.4715	66.032	67.668
	1 B ₁	5.18	0.005		-0.054
	1 B ₂	5.28	0.028		0.163
	1 A ₁	6.71	0.003		-0.042
	2 A ₁	6.79	0.027		-0.365
	2 B ₁	7.27	0.041		-1.547
experiment ^{99,63}	2 B ₂	7.34	0.008		-2.300
	B ₁	4.41	n.a.		-0.0002
	B ₂	4.96	0.04/0.041		0.1/0.15

^a Comparison of B3LYP, CAMB3LYP and CCSD results. The molecule lies on the yz plane, with the C_2 symmetry axis along z .

(and third) band. Only two bands appear within the frequency range considered in the simulated experimental

spectrum.⁹⁹ Comparing our DFT results with those in ref 42, we note that none of the two hybrid functionals we

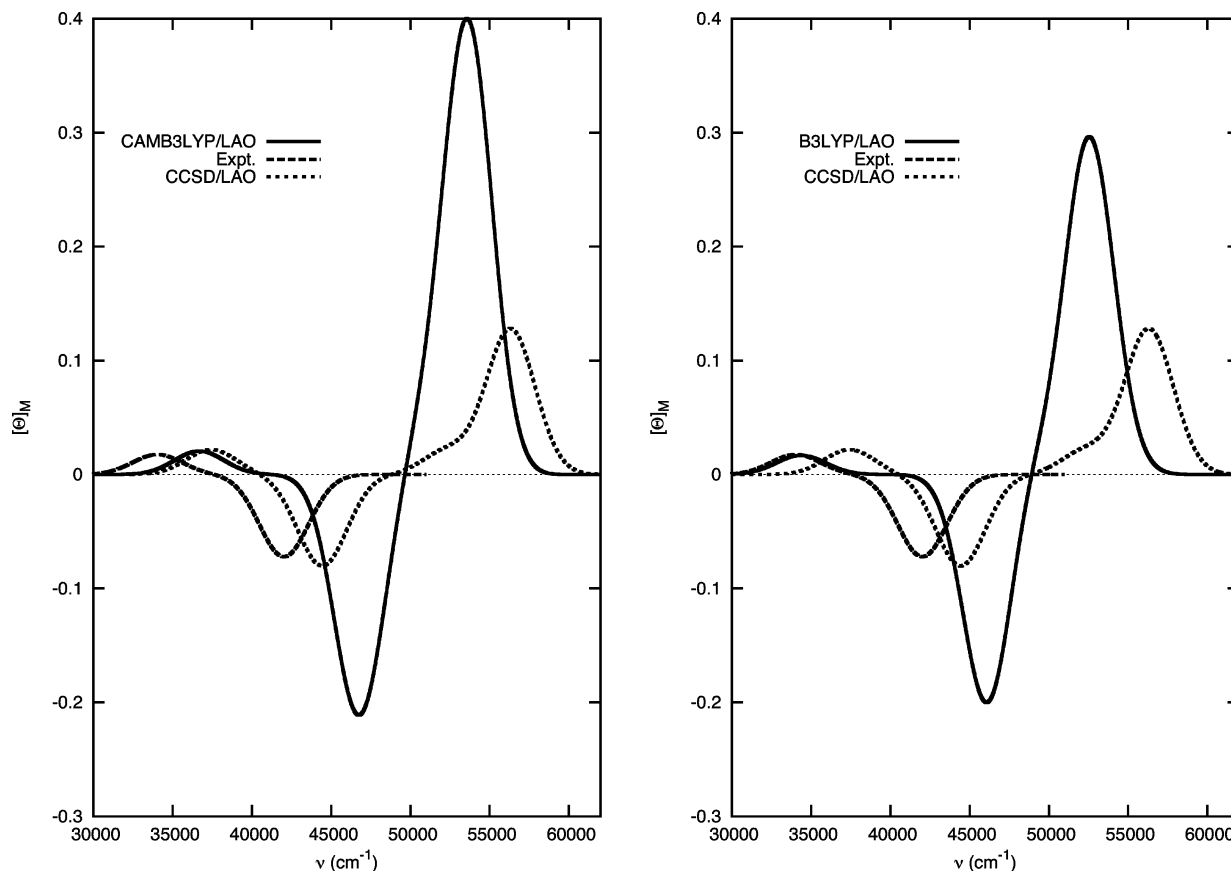


Figure 6. Pyrimidine: Comparison of simulated MCD spectra for B3LYP, CAMB3LYP, and CCSD.

have considered yields a second $n \rightarrow \pi^*$ transition of B_1 symmetry in the range 37 000–45 000 cm^{-1} as obtained instead by Seth et al. using the SAOP functional and the QZ3PID basis set.

3.2.2. Phosphabenzene. Turning our attention to phosphabenzene, we recall that in the experimental absorption spectrum the first two transitions (assigned as B_2 and B_1) only appear as shoulders on the band for the strong transition (of A_1 symmetry) peaked at around 5.21 eV and characterized by an oscillator strength f of 0.16. The same symmetry classification is reproduced by B3LYP and CAMB3LYP (as well as CCSD), but not by Hartree–Fock and LDA. Both hybrid functionals yield \mathcal{B} terms for the first three dipole-allowed transitions of the same signs as those computed by CCSD and those obtained from the experimental measurement, but again the magnitude of the terms varies significantly from method to method. This, together with the different spacing in the frequency of the various excitations yielded by the two different functionals, strongly affects the appearance of the simulated spectra reproduced in Figure 7, because of cancellation effects between bands associated with \mathcal{B} terms of opposite sign falling relatively close to each other. The most remarkable consequence is the disappearance, in the B3LYP case, of the positive peak at around 40 000–42 000 cm^{-1} . As in the pyrimidine case, an estimate of the experimental spectrum from reference 100 has been included, obtained by convolution of Gaussian line shape functions built from the experimentally derived \mathcal{B} terms of Waluk et al.¹⁰⁰

3.2.3. Pyridine. Only B3LYP and CAMB3LYP results are reported for pyridine, together with the CCSD results from ref 36, see Table 10. Also for this molecule our previous study at the CCSD level revealed strong cancellation effects between oppositely signed MCD bands, see ref 36. Similar to what was observed for pyrimidine and phosphabenzene, the two hybrid functionals reproduce the CCSD symmetry assignment for the first four transitions (consistent with the experimental assignment for the first two transitions). Like CCSD, the two functionals also yield a weaker negative \mathcal{B} terms for the first transition and a stronger positive one for the second one, whereas differences are seen concerning the sign and relative intensity of the next two transitions, of A_1 symmetry. Simulating the MCD spectrum as done for the other two benzene analogues, see Figure 8, shows once again how the relative position of the excitation energy and the magnitude of the \mathcal{B} terms oppositely signed dramatically affects the overall spectrum: the first positive MCD band, so weak to disappear completely in the picture at the CCSD level (in accordance to the experimental findings), is strongly enhanced at the DFT level because of the larger spacing between the first two transitions, which reduces the cancellation effect occurring between the weaker negative \mathcal{B} of the first transition and the much larger positive \mathcal{B} of the second transition. Similarly the second, negative, band is enhanced and the third, positive, band that dominates the DFT spectra falls at much lower frequencies than the analogous band in the CCSD spectrum. Our simulated MCD spectrum is qualitatively similar to the

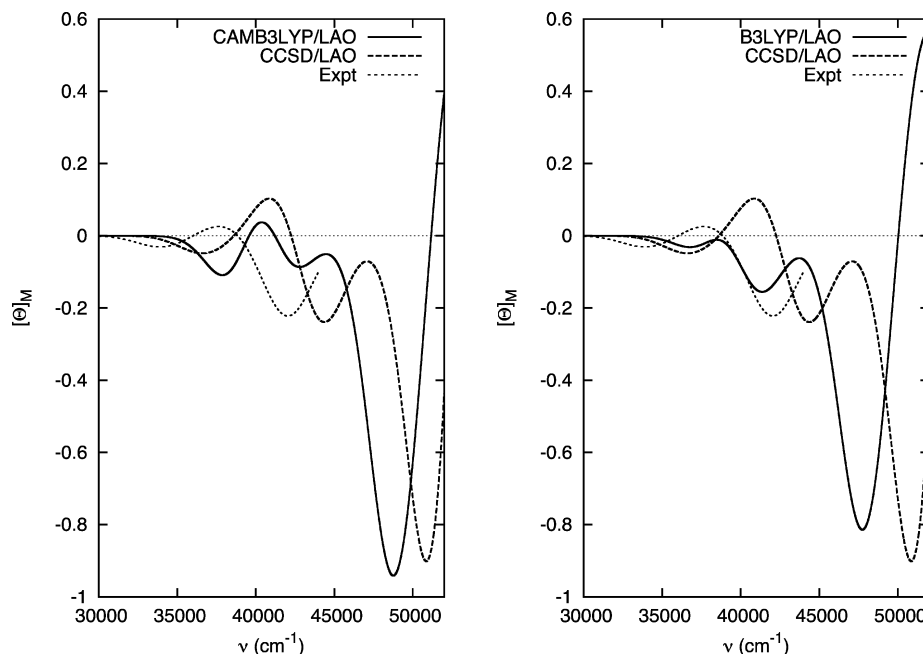


Figure 7. Phosphabenzene: Comparison of simulated MCD spectra for B3LYP, CAMB3LYP, and CCSD.

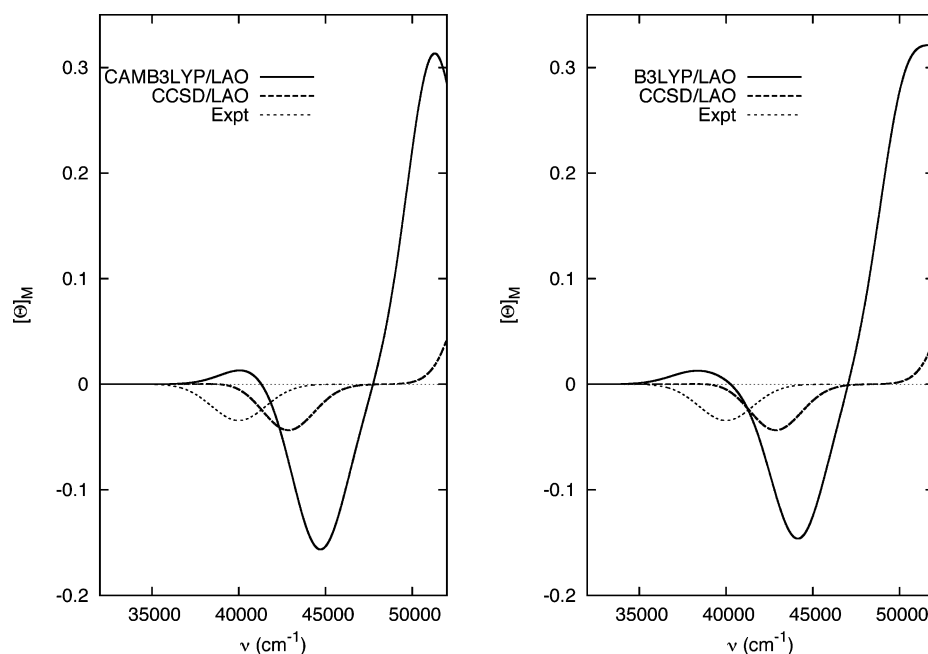


Figure 8. Pyridine: MCD spectra simulations. B3LYP/CAMB3LYP.

one obtained by Seth and co-workers in ref 42 in the frequency range of the first two dipole allowed transitions.

4. Summary and Conclusions

We have presented a Lagrangian approach to obtain gauge-origin independent expressions for the Verdet constant $V(\omega)$ of the Faraday effect and the \mathcal{B} term of magnetic circular dichroism at the Hartree–Fock and TD-DFT levels of theory in an atomic-orbital based formalism suitable for linear-scaling. An important feature of the formulation is a projection to the orthogonal complement of the excited reference state to remove unphysical divergences in the response equations solved for the \mathcal{B} term of MCD, and its implementation in terms of a modified preconditioned

conjugated gradient method for the computation of the projected property Lagrange multipliers of the transition-moment derivative.

The methodology has been used to compute the Verdet constant of H_2 , HF, CO, N_2O , and $\text{CH}_3\text{CH}_2\text{CH}_3$, as well as the \mathcal{B} terms and corresponding MCD spectra of pyrimidine, phosphabenzene, and pyridine. The results have been benchmarked against CCSD values obtained applying a gauge-origin independent method previously developed in our group.¹⁷

Our basis set study of the Verdet constant shows that, for the functionals and molecules here considered, the model-inherent error is much larger than the basis set error, and the smallest singly augmented double- ζ basis set would

therefore suffice for qualitative comparison with experiment. Using LAOs reduces the basis set error by between 60% and 90% for singly augmented and between 0% and 50% for doubly augmented basis sets. We expect this reduction to be even greater for larger, unsymmetric molecules, where there is no obvious choice of gauge origin.

Both magneto-optical properties appear to represent quite a challenging task for both the Hartree–Fock and Kohn–Sham methods. The \mathcal{B} term of MCD in particular proved very challenging, since it requires an accurate description of the excited states, the transition and the magnetic perturbation. For systems with well separated transitions the TD-DFT MCD spectra reproduce most of the qualitative features of the CCSD spectra, but information is often lost in the simulation of the MCD spectrum, especially in presence of strong overlapping bands, and the calculated \mathcal{B} terms would then appear to be more useful. For the individual \mathcal{B} terms, the deviation from experiment seems to lack a systematic behavior for both valence and Rydberg states, even for the CAMB3LYP functional, which has an improved long-range behavior and has been shown to better describe, compared to other functionals, excited state energies and excitations to states of Rydberg character.¹⁰¹ We believe that an extensive benchmark study is required before a specific DFT functional can be recommended. In this perspective the gauge-origin-independent CCSD method of ref 17 may prove an invaluable tool for benchmarking the \mathcal{B} term of MCD and determining the reliability of the \mathcal{B} terms yielded by different DFT functionals. Work is in progress in this direction.

Acknowledgment. This work has received support from the Lundbeck Foundation and the Danish Center for Scientific computing. S.C. acknowledges her affiliation to the Istituto Nazionale Scienze e Tecnologia dei Materiali (INSTM) and financial support from the Italian Ministero dell’Università e Ricerca (MiUR), Programmi di ricerca di interesse nazionale, PRIN2006. A.J.T. has received support from the Norwegian Research Council (Grant No. 177558/V30).

Appendix

Appendix A: Modified Preconditioned Conjugated Gradient Method for the Computation of the Projected Property Lagrange Multipliers of the Transition-Moment Derivative. The variational condition of the \mathcal{B} -term Lagrangian with respect to the excited state vector yields a linear response equation

$$\mathbf{P}^{\dagger}(\mathbf{E}^{[2]} - \omega_{\rho}\mathbf{S}^{[2]})\mathbf{P}\tilde{\lambda} = \mathbf{P}^{\dagger}\mathbf{A}^{[1]} \quad (74)$$

where the solution vector must be orthogonal to the excitation vector using an $\mathbf{S}^{[2]}$ metric, i.e. $\mathbf{b}^{i\dagger}\mathbf{S}^{[2]}\mathbf{b}^j = \delta_{ij}$ (or $-\delta_{ij}$ for the deexcitation vectors \mathbf{b}^{-i}). The projector which ensures this orthonormality \mathbf{P}^f is given in eq 63.

To solve eq 74, we would like to use the response solver of ref 74, which is based on an iterative preconditioned conjugate gradient algorithm (using a reduced space). Moreover the method always implicitly add trial vectors in

pairs, a strategy that, for the eigenvalue problem, guarantees real eigenvalues and fast convergence.

Straightforward application of the projector \mathbf{P}^f on the trial vector does not comply with the use of an iterative algorithm where trial vectors are added in pairs, since \mathbf{b}^{-f} has no paired counterpart and a non-vanishing component in the solution. Thus special considerations has to be given to the \mathbf{b}^{-f} vector when an algorithm is used where trial vectors are added in pairs.

The iterative procedure given in section II.A of ref 74 therefore has to be slightly modified when solving eq 74. The modifications may be summarized as follows:

- Instead of a basis $2n$ trial vectors (i.e., n pairs), the procedure uses a $2 + 2n$ trial vector basis, where the first two vectors are always chosen as the excitation vector \mathbf{b}^f , and its de-excitation counterpart \mathbf{b}^{-f} (\mathbf{b}^f and \mathbf{b}^{-f} are the paired vectors obtained solving the eigenvalue equation, with eigenvalues ω_f and $-\omega_f$, respectively). The remaining $2n$ vectors (or n pairs) are generated (at each iteration) as in the standard procedure⁷⁴ but with the additional requirement that they are always kept orthogonal, with respect to $\mathbf{S}^{[2]}$, to both \mathbf{b}^f and \mathbf{b}^{-f}

$$\mathbf{b}_i = \mathbf{P}^{f,-f}(\mathbf{b}_i); \quad \forall i = 3, 2n + 2 \quad (75)$$

The projection matrix

$$\mathbf{P}^{f,-f} = \mathbf{I} - \mathbf{b}^f\mathbf{b}^{f\dagger}\mathbf{S}^{[2]} + \mathbf{b}^{-f}\mathbf{b}^{-f\dagger}\mathbf{S}^{[2]} \quad (76)$$

ensures this orthogonality. The $\{\mathbf{b}_i, i = 3, \dots, 2n + 2\}$ trial vectors are, as previously, normalized and orthogonalized against each other in a standard Euclidian way, but not against the first two vectors, $\mathbf{b}_1 = \mathbf{b}^f$ and $\mathbf{b}_2 = \mathbf{b}^{-f}$.

- The above $2n + 2$ trial basis and the corresponding linear transformations σ^{2n+2} and ρ^{2n+2} are used to set up the response equation in a reduced space of dimension $2n + 2$. Because of the above choice of trial basis, however, the reduced space equation

$$({}^R\mathbf{E}^{[2]} - {}^R\omega_f\mathbf{S}^{[2]}){}^R\mathbf{X} = {}^R\mathbf{A}^{[1]} \quad (77)$$

$${}^R\mathbf{X} = {}^R\mathbf{A}^{[1]} \quad (78)$$

will have the following “blocked” structure

$$\begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & \mathcal{R}_{2,2} & \dots & \mathcal{R}_{2,2n+2} \\ \vdots & \vdots & \dots & \vdots \\ 0 & \mathcal{R}_{2n+2,2} & \dots & \mathcal{R}_{2n+2,2n+2} \end{pmatrix} \begin{pmatrix} {}^R\mathbf{X}_1 \\ {}^R\mathbf{X}_2 \\ \vdots \\ {}^R\mathbf{X}_{2n+2} \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbf{b}^{-f\dagger}\mathbf{P}^{\dagger}\mathbf{A}^{[1]} \\ \vdots \\ \mathbf{b}_{2n+2}^{\dagger}\mathbf{P}^{\dagger}\mathbf{A}^{[1]} \end{pmatrix} \quad (79)$$

The non-zero subblock of the above equation gives the reduced space solution of the projected equation, whereas the element ${}^R\mathbf{X}_1$ (which would give the component of the solution vector along the \mathbf{b}^f excited state vector) is manually set to zero.

- The solution of the reduced space is used to construct the current solution vector (at each iteration) \mathbf{X} by expansion onto the trial basis

$$X = \sum_{i=2}^{2n+2} {}^R X_i \mathbf{b}_i \quad (80)$$

that is, with no contribution from \mathbf{b}^f but with contribution from \mathbf{b}^{-f} . The residual is then obtained in the usual way

$$\mathbf{R} = \mathbf{E}^{[2]} X - \omega_f \mathbf{S}^{[2]} X - \mathbf{P}^{ff} \mathbf{A}^{[1]} \quad (81)$$

$$\mathbf{R} = \sum_{i=2}^{2n} {}^R X_i (\boldsymbol{\sigma}_i - \omega_f \boldsymbol{\rho}_i) - \mathbf{P}^{ff} \mathbf{A}^{[1]} \quad (82)$$

and preconditioned according to

$$\mathbf{R}_p = \mathbf{M}^{-1} \mathbf{R} \quad (83)$$

where the preconditioner \mathbf{M} is an appropriately chosen matrix. For further information on how \mathbf{M} is chosen, see ref 74.

- The preconditioned residual is projected with respect to the $\{\mathbf{b}^f, \mathbf{b}^{-f}\}$ space as $(\mathbf{P}^{f-f} \mathbf{R}_p)$, and a new pair of trial vectors is then obtained according to the standard procedure of ref 74 (i.e. by normalization and orthogonalization versus the previous $\{\mathbf{b}_3, \dots, \mathbf{b}_{2n+2}\}$ trial vectors). The new pair of trial vectors is added to the original trial basis, the reduced space is enlarged and the iterative procedure continued until convergence, that is, until the norm of the residual is smaller than a given threshold.

Appendix B: Exchange-Correlation Contribution to the Derivative Quantities. The determination of gauge-origin independent analytic expressions for the MOA properties requires that both the derivative of the exchange-correlation contribution to the Kohn-Sham matrix $(\partial \mathbf{F}^{\text{xc}}) / (\partial \mathbf{B})|_{\mathbf{B}=0}$ and the derivative of the exchange correlation contribution to the (generalized) Kohn-Sham Hessian $(\partial \mathbf{G}^{\text{xc}}) / (\partial \mathbf{B})|_{\mathbf{B}=0}$ are evaluated. The object of this Appendix is to derive equations for both these exchange-correlation derivative contributions.

B.1. Exchange-Correlation Contribution to the Derivative Kohn–Sham Matrix. The exchange-correlation energy requires the integration over all space of some functional $f = f[\rho, \nabla \rho]$ which depends on the electron density $\rho = \rho(\mathbf{r})$ and on the gradient of the density $\nabla \rho = \nabla \rho(\mathbf{r})$

$$E_{\text{xc}} = \int f[\rho, \nabla \rho] \mathbf{d}\mathbf{r} \quad (84)$$

The exchange-correlation contribution to the Kohn–Sham matrix is given by

$$F_{\mu\nu}^{\text{xc}} = \int v_{\text{xc}} \Omega_{\mu\nu}(\mathbf{r}) \mathbf{d}\mathbf{r} \quad (85)$$

where v_{xc} is the exchange correlation potential given in eq 18. Using the form of E_{xc} in eq 84 gives^{102,103}

$$v_{\text{xc}} = \frac{\delta E_{\text{xc}}}{\delta \rho} = \frac{\partial f[\rho, \nabla \rho]}{\partial \rho} - \nabla \frac{\partial f[\rho, \nabla \rho]}{\partial \nabla \rho} \quad (86)$$

To simplify the evaluation of eq 85, we insert eq 86 and carry out a partial integration assuming that the constant term vanishes because f is local

$$F_{\mu\nu}^{\text{xc}} = \int \frac{\partial f}{\partial \rho} \Omega_{\mu\nu}(\mathbf{r}) \mathbf{d}\mathbf{r} + \int \frac{\partial f}{\partial \nabla \rho} \nabla \Omega_{\mu\nu}(\mathbf{r}) \mathbf{d}\mathbf{r} \quad (87)$$

Considering $\xi = |\nabla \rho|$ as the fundamental variable and applying the chain rule gives

$$F_{\mu\nu}^{\text{xc}} = \int \frac{\partial f}{\partial \rho} \Omega_{\mu\nu}(\mathbf{r}) \mathbf{d}\mathbf{r} + \int \frac{\partial f}{\partial \xi} \frac{\partial \xi}{\partial \nabla \rho} \nabla \Omega_{\mu\nu}(\mathbf{r}) \mathbf{d}\mathbf{r} \quad (88)$$

From the definition of ξ , we obtain

$$\frac{\partial \xi}{\partial \nabla \rho(\mathbf{r})} = \frac{\nabla \rho(\mathbf{r})}{\xi} \quad (89)$$

which may be inserted into eq 88 giving

$$F_{\mu\nu}^{\text{xc}} = \int \frac{\partial f}{\partial \rho} \Omega_{\mu\nu}(\mathbf{r}) \mathbf{d}\mathbf{r} + \int \frac{\partial f}{\partial \xi} \frac{\nabla \rho(\mathbf{r})}{\xi} \nabla \Omega_{\mu\nu} \mathbf{d}\mathbf{r} \quad (90)$$

The magnetic field dependence of the Kohn–Sham matrix contribution may be introduced by replacing the atomic orbitals $\chi_{\mu}(\mathbf{r})$ with the London orbitals $\omega_{\mu}(\mathbf{B}, \mathbf{r})$ of eq 29. The electron density then depends explicitly on the magnetic field as

$$\rho(\mathbf{r}, \mathbf{B}) = \sum_{\mu\nu} \exp\left(\frac{i}{2} \mathbf{B} \cdot \mathbf{R}_{MN} \times \mathbf{r}\right) \chi_{\mu}^*(\mathbf{r}) \chi_{\nu}(\mathbf{r}) D_{\nu\mu} = \sum_{\mu\nu} \tilde{\Omega}_{\mu\nu}(\mathbf{B}, \mathbf{r}) D_{\nu\mu} \quad (91)$$

and the exchange-correlation contribution to the Kohn–Sham matrix becomes

$$F_{\mu\nu}^{\text{xc}} = \int \frac{\partial \tilde{f}}{\partial \rho} \tilde{\Omega}_{\mu\nu} \mathbf{d}\mathbf{r} + \int \frac{\partial \tilde{f}}{\partial \xi} \frac{\nabla \tilde{\rho}(\mathbf{r})}{\xi} \nabla \tilde{\Omega}_{\mu\nu} \mathbf{d}\mathbf{r} \quad (92)$$

where tilde is used to denote the explicit magnetic field dependence. The first derivative of the density and of the gradient of the density with respect to the field at zero field is zero⁸⁵

$$\left. \frac{\partial \rho(\mathbf{r}, \mathbf{B})}{\partial \mathbf{B}} \right|_{\mathbf{B}=0} = 0 \quad (93)$$

$$\left. \frac{\partial \nabla \rho(\mathbf{r}, \mathbf{B})}{\partial \mathbf{B}} \right|_{\mathbf{B}=0} = 0 \quad (94)$$

This simplifies the differentiation of the exchange-correlation contribution to the Kohn–Sham matrix yielding

$$\left. \frac{\partial F_{\mu\nu}^{\text{xc}}}{\partial \mathbf{B}} \right|_{\mathbf{B}=0} = \int \frac{\partial f}{\partial \rho} \frac{\partial \tilde{\Omega}_{\mu\nu}}{\partial \mathbf{B}} \Big|_{\mathbf{B}=0} \mathbf{d}\mathbf{r} + \int \frac{\partial f}{\partial \xi} \frac{\nabla \rho}{\xi} \frac{\partial \nabla \tilde{\Omega}_{\mu\nu}}{\partial \mathbf{B}} \Big|_{\mathbf{B}=0} \mathbf{d}\mathbf{r} \quad (95)$$

From the definition of the overlap distribution $\tilde{\Omega}_{\mu\nu}$, we obtain

$$\left. \frac{\partial \tilde{\Omega}_{\mu\nu}}{\partial \mathbf{B}} \right|_{\mathbf{B}=0} = \frac{i}{2} (\mathbf{R}_{MN} \times \mathbf{r}) \Omega_{\mu\nu} \quad (96)$$

The first term in eq 95 therefore becomes

$$\int \frac{\partial f}{\partial \rho} \frac{\partial \tilde{\Omega}_{\mu\nu}}{\partial \mathbf{B}} \Big|_{\mathbf{B}=0} \mathbf{d}\mathbf{r} = \frac{i}{2} \int (\mathbf{R}_{MN} \times \mathbf{r}) \frac{\partial f}{\partial \rho} \Omega_{\mu\nu} \mathbf{d}\mathbf{r} \quad (97)$$

To determine the second term we derive

$$\frac{\partial \omega_q^*}{\partial r_c} = \frac{\partial}{\partial r_c} \exp \left\{ \frac{i}{2} \sum_{ijk} \varepsilon_{ijk} B_i(\mathbf{R}_{QO})_j r_k \right\} \chi_q^* \quad (98)$$

$$\begin{aligned} \frac{\partial \omega_q^*}{\partial r_c} = & \frac{i}{2} \left(\sum_{ij} \varepsilon_{ijc} B_i(\mathbf{R}_{QO})_j \right) \exp \left\{ \frac{i}{2} \sum_{ijk} \varepsilon_{ijk} B_i(\mathbf{R}_{QO})_j r_k \right\} \chi_q^* + \\ & \exp \left\{ \frac{i}{2} \sum_{ijk} \varepsilon_{ijk} B_i(\mathbf{R}_{QO})_j r_k \right\} \frac{\partial \chi_q^*}{\partial r_c} \end{aligned} \quad (99)$$

where the Levi–Civita tensor has been introduced to describe the cross-product and c denotes a spatial electronic coordinate component. A subsequent differentiation with respect to a magnetic field yields

$$\left. \frac{\partial}{\partial B_\gamma} \frac{\partial \omega_q^*}{\partial r_c} \right|_{\mathbf{B}=0} = \frac{i}{2} \left(\sum_j \varepsilon_{\gamma jc} (\mathbf{R}_{QO})_j \right) \chi_q + \frac{i}{2} (\mathbf{R}_{QO} \times \mathbf{r})_\gamma \frac{\partial \chi_q}{\partial r_c} \quad (100)$$

which may be used to obtain

$$\left. \frac{\partial}{\partial B_\gamma} \frac{\partial \tilde{\Omega}_{\mu\nu}}{\partial r_c} \right|_{\mathbf{B}=0} = \frac{i}{2} \left(\sum_j \varepsilon_{\gamma jc} (\mathbf{R}_{MN})_j \right) \Omega_{\mu\nu} + \frac{i}{2} (\mathbf{R}_{MN} \times \mathbf{r})_\gamma \frac{\partial \Omega_{\mu\nu}}{\partial r_c} \quad (101)$$

Inserting eq 101 into the second term of eq 95 gives

$$\begin{aligned} & \int \frac{\partial f}{\partial \xi} \frac{\partial \nabla \rho}{\xi} \frac{\partial \nabla \tilde{\Omega}_{\mu\nu}}{\partial B_\gamma} \Big|_{\mathbf{B}=0} \mathbf{dr} \\ &= \int \frac{\partial f}{\partial \xi} \frac{1}{\xi} \sum_c \frac{\partial \rho}{\partial r_c} \frac{\partial}{\partial B_\gamma} \frac{\partial \tilde{\Omega}_{\mu\nu}}{\partial r_c} \Big|_{\mathbf{B}=0} \mathbf{dr} \\ &= \int \frac{\partial f}{\partial \xi} \frac{1}{\xi} \sum_{cj} \frac{\partial \rho}{\partial r_c} \frac{i}{2} (\varepsilon_{\gamma jc} (\mathbf{R}_{MN})_j) \Omega_{\mu\nu} \mathbf{dr} \\ &+ \int \frac{\partial f}{\partial \xi} \frac{1}{\xi} \sum_c \frac{i}{2} (\mathbf{R}_{MN} \times \mathbf{r})_\gamma \frac{\partial \rho}{\partial r_c} \frac{\partial \Omega_{\mu\nu}}{\partial r_c} \mathbf{dr} \\ &= \int \frac{\partial f}{\partial \xi} \frac{1}{\xi} \frac{i}{2} ((\mathbf{R}_{MN} \times \nabla \rho)_\gamma \Omega_{\mu\nu} + (\mathbf{R}_{MN} \times \mathbf{r})_\gamma \nabla \rho \nabla \Omega_{\mu\nu}) \mathbf{dr} \end{aligned} \quad (102)$$

The exchange-correlation contribution to the derivative Kohn–Sham matrix may therefore be obtained as

$$\begin{aligned} \left. \frac{\partial F_{\mu\nu}^{\text{xc}}}{\partial B} \right|_{\mathbf{B}=0} = & \frac{i}{2} \int (\mathbf{R}_{MN} \times \mathbf{r}) \left[\frac{\partial f}{\partial \rho(\mathbf{r})} \Omega_{\mu\nu}(\mathbf{r}) + \right. \\ & \left. \frac{\partial f}{\partial \xi} \frac{\nabla \rho(\mathbf{r})}{\xi} \nabla \Omega_{\mu\nu}(\mathbf{r}) \right] \mathbf{dr} + \frac{i}{2} \int (\mathbf{R}_{MN} \times \frac{\nabla \rho(\mathbf{r})}{\xi}) \frac{\partial f}{\partial \xi} \Omega_{\mu\nu}(\mathbf{r}) \mathbf{dr} \end{aligned} \quad (103)$$

where we have used eq 97 and eq 102.

B.2. Exchange-Correlation Contribution to the Derivative Generalized Kohn–Sham Hessian Matrix. The exchange-correlation contribution to the (generalized) Kohn–Sham Hessian is determined by⁷⁵

$$G_{\mu\nu}^{\text{xc}}(\mathbf{M}) = \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\delta v_{\text{xc}}(\mathbf{r})}{\delta \rho(\mathbf{s})} \Omega_{\mu\nu}(\mathbf{r}) \Omega_{\rho\sigma}(\mathbf{s}) \mathbf{drds} \quad (104)$$

where \mathbf{M} is a general ‘‘perturbed’’ density matrix, in our case

either \mathbf{D}^b or \mathbf{D}^f . Using eq 86, we obtain

$$\begin{aligned} G_{\mu\nu}^{\text{xc}}(\mathbf{M}) = & \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\delta}{\delta \rho(\mathbf{s})} \frac{\partial f[\rho(\mathbf{r}), \nabla \rho(\mathbf{r})]}{\partial \rho(\mathbf{r})} \Omega_{\mu\nu}(\mathbf{r}) \Omega_{\rho\sigma}(\mathbf{s}) \mathbf{drds} - \\ & \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\delta}{\delta \rho(\mathbf{s})} \nabla_r \left(\frac{\partial f[\rho(\mathbf{r}), \nabla \rho(\mathbf{r})]}{\partial \nabla \rho(\mathbf{r})} \right) \Omega_{\mu\nu}(\mathbf{r}) \Omega_{\rho\sigma}(\mathbf{s}) \mathbf{drds} \end{aligned} \quad (105)$$

where the subscript ∇_r denotes the variable with respect to which we differentiate. Using partial integration on the second term and assuming that the constant term vanishes because f is local gives

$$\begin{aligned} G_{\mu\nu}^{\text{xc}}(\mathbf{M}) = & \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\delta}{\delta \rho(\mathbf{s})} \frac{\partial f[\rho(\mathbf{r}), \nabla \rho(\mathbf{r})]}{\partial \rho(\mathbf{r})} \Omega_{\mu\nu}(\mathbf{r}) \Omega_{\rho\sigma}(\mathbf{s}) \mathbf{drds} + \\ & \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\delta}{\delta \rho(\mathbf{s})} \frac{\partial f[\rho(\mathbf{r}), \nabla \rho(\mathbf{r})]}{\partial \nabla \rho(\mathbf{r})} \nabla_r (\Omega_{\mu\nu}(\mathbf{r}) \Omega_{\rho\sigma}(\mathbf{s})) \mathbf{drds} \end{aligned} \quad (106)$$

Using the chain rule for functional derivatives

$$\frac{\delta F}{\delta g(\mathbf{y})} = \int \frac{\delta F}{\delta f(\mathbf{x})} \frac{\delta f(\mathbf{x})}{\delta g(\mathbf{y})} \mathbf{dx} \quad (107)$$

yields

$$G_{\mu\nu}^{\text{xc}}(\mathbf{M}) = P_1 + P_2 + P_3 + P_4 \quad (108)$$

$$P_1 = \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\delta \partial f[\rho(\mathbf{r}), \nabla \rho(\mathbf{r})]}{\delta \rho(\mathbf{t}) \partial \rho(\mathbf{r})} \frac{\delta \rho(\mathbf{t})}{\delta \rho(\mathbf{s})} \Omega_{\mu\nu}(\mathbf{r}) \Omega_{\rho\sigma}(\mathbf{s}) \mathbf{drdsdt} \quad (109)$$

$$P_2 = \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\delta \partial f[\rho(\mathbf{r}), \nabla \rho(\mathbf{r})]}{\delta \rho(\mathbf{t}) \partial \nabla \rho(\mathbf{r})} \times \frac{\delta \rho(\mathbf{t})}{\delta \rho(\mathbf{s})} \nabla_r (\Omega_{\mu\nu}(\mathbf{r}) \Omega_{\rho\sigma}(\mathbf{s})) \mathbf{drdsdt} \quad (110)$$

$$P_3 = \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\delta \partial f[\rho(\mathbf{r}), \nabla \rho(\mathbf{r})]}{\delta \nabla \rho(\mathbf{t}) \partial \rho(\mathbf{r})} \times \frac{\delta \nabla \rho(\mathbf{t})}{\delta \rho(\mathbf{s})} \Omega_{\mu\nu}(\mathbf{r}) \Omega_{\rho\sigma}(\mathbf{s}) \mathbf{drdsdt} \quad (111)$$

$$P_4 = \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\delta \partial f[\rho(\mathbf{r}), \nabla \rho(\mathbf{r})]}{\delta \nabla \rho(\mathbf{t}) \partial \nabla \rho(\mathbf{r})} \times \frac{\delta \nabla \rho(\mathbf{t})}{\delta \rho(\mathbf{s})} \nabla_r (\Omega_{\mu\nu}(\mathbf{r}) \Omega_{\rho\sigma}(\mathbf{s})) \mathbf{drdsdt} \quad (112)$$

Using the relation between functional and standard derivatives for composite functions

$$\frac{\delta \rho(\mathbf{t})}{\delta \rho(\mathbf{s})} = \delta(\mathbf{t} - \mathbf{s}) \quad (113)$$

$$\frac{\delta f[\rho(\mathbf{r})]}{\delta \rho(\mathbf{t})} = \frac{\partial f[\rho(\mathbf{r})]}{\partial \rho(\mathbf{r})} \delta(\mathbf{r} - \mathbf{t}) \quad (114)$$

and integrating first over \mathbf{s} and then \mathbf{t} , the first two terms of eq 108 may be rewritten

$$\begin{aligned} P_1 + P_2 = & \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\partial^2 f}{\partial \rho^2} \Omega_{\mu\nu} \Omega_{\rho\sigma} \mathbf{dr} + \\ & \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\partial^2 f}{\partial \rho \partial \nabla \rho} \nabla (\Omega_{\mu\nu}) \Omega_{\rho\sigma} \mathbf{dr} \end{aligned} \quad (115)$$

Using the relation

$$\frac{\delta \nabla \rho(\mathbf{t})}{\delta \rho(\mathbf{s})} = \nabla_t \frac{\delta \rho(\mathbf{t})}{\delta \rho(\mathbf{s})} = \nabla_t \delta(\mathbf{t} - \mathbf{s}) \quad (116)$$

the last two terms become

$$P_3 + P_4 = \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\delta \partial f[\rho(\mathbf{r}), \nabla \rho(\mathbf{r})]}{\delta \nabla \rho(\mathbf{t}) \partial \rho(\mathbf{r})} \times \\ \nabla_t \delta(\mathbf{t} - \mathbf{s}) \Omega_{\mu\nu}(\mathbf{r}) \Omega_{\rho\sigma}(\mathbf{s}) d\mathbf{r} d\mathbf{s} + \\ \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\delta \partial f[\rho(\mathbf{r}), \nabla \rho(\mathbf{r})]}{\delta \nabla \rho(\mathbf{t}) \partial \nabla \rho(\mathbf{r})} \times \\ \nabla_t \delta(\mathbf{t} - \mathbf{s}) \nabla_r (\Omega_{\mu\nu}(\mathbf{r})) \Omega_{\rho\sigma}(\mathbf{s}) d\mathbf{r} d\mathbf{s} \quad (117)$$

Performing a partial integration, assuming that the constant term vanishes because f is a local function, gives

$$P_3 + P_4 = - \sum_{\rho\sigma} M_{\sigma\rho} \int \nabla_t \left(\frac{\delta \partial f[\rho(\mathbf{r}), \nabla \rho(\mathbf{r})]}{\delta \nabla \rho(\mathbf{t}) \partial \rho(\mathbf{r})} \right) \times \\ \delta(\mathbf{t} - \mathbf{s}) \Omega_{\mu\nu}(\mathbf{r}) \Omega_{\rho\sigma}(\mathbf{s}) d\mathbf{r} d\mathbf{s} dt - \\ \sum_{\rho\sigma} M_{\sigma\rho} \int \nabla_t \left(\frac{\delta \partial f[\rho(\mathbf{r}), \nabla \rho(\mathbf{r})]}{\delta \nabla \rho(\mathbf{t}) \partial \nabla \rho(\mathbf{r})} \right) \times \\ \delta(\mathbf{t} - \mathbf{s}) \nabla_r (\Omega_{\mu\nu}(\mathbf{r})) \Omega_{\rho\sigma}(\mathbf{s}) d\mathbf{r} d\mathbf{s} dt \quad (118)$$

Performing the integration over \mathbf{t} yields

$$P_3 + P_4 = - \sum_{\rho\sigma} M_{\sigma\rho} \int \nabla_s \left(\frac{\delta \partial f[\rho(\mathbf{r}), \nabla \rho(\mathbf{r})]}{\delta \nabla \rho(\mathbf{s}) \partial \rho(\mathbf{r})} \right) \times \\ \Omega_{\mu\nu}(\mathbf{r}) \Omega_{\rho\sigma}(\mathbf{s}) d\mathbf{r} d\mathbf{s} - \sum_{\rho\sigma} M_{\sigma\rho} \int \nabla_s \left(\frac{\delta \partial f[\rho(\mathbf{r}), \nabla \rho(\mathbf{r})]}{\delta \nabla \rho(\mathbf{s}) \partial \nabla \rho(\mathbf{r})} \right) \times \\ \nabla_r (\Omega_{\mu\nu}(\mathbf{r})) \Omega_{\rho\sigma}(\mathbf{s}) d\mathbf{r} d\mathbf{s} \quad (119)$$

Yet another partial integration and an integration over \mathbf{s} using eq 114 yields

$$P_3 + P_4 = \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\partial^2 f}{\partial \rho \partial \nabla \rho} \Omega_{\mu\nu} \nabla \Omega_{\rho\sigma} d\mathbf{r} + \\ \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\partial^2 f}{\partial (\nabla \rho)^2} \nabla \Omega_{\mu\nu} \nabla \Omega_{\rho\sigma} d\mathbf{r} \quad (120)$$

The total $G_{\mu\nu}^{\text{xc}}(\mathbf{M})$ may be obtained from eq 115 and 120 giving

$$G_{\mu\nu}^{\text{xc}}(\mathbf{M}) = \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\partial^2 f}{\partial \rho^2} \Omega_{\mu\nu} \Omega_{\rho\sigma} + \frac{\partial^2 f}{\partial \rho \partial \nabla \rho} (\Omega_{\mu\nu} \nabla \Omega_{\rho\sigma} + \\ \Omega_{\rho\sigma} \nabla \Omega_{\mu\nu}) d\mathbf{r} + \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\partial^2 f}{\partial (\nabla \rho)^2} \nabla \Omega_{\mu\nu} \nabla \Omega_{\rho\sigma} d\mathbf{r} \quad (121)$$

The general chain rule

$$\frac{\partial f}{\partial \nabla \rho} = \frac{\partial f}{\partial \xi} \frac{\partial \xi}{\partial \nabla \rho} = \frac{\nabla \rho}{\xi} \quad (122)$$

$$\frac{\partial^2 f}{\partial (\nabla \rho)^2} = \frac{\partial^2 f}{\partial \xi^2} \left(\frac{\partial \xi}{\partial \nabla \rho} \right)^2 + \frac{\partial f}{\partial \xi} \frac{\partial^2 \xi}{\partial (\nabla \rho)^2} = \frac{\partial^2 f}{\partial \xi^2} \left(\frac{\nabla \rho}{\xi} \right)^2 \quad (123)$$

can now be applied in order to introduce $\xi = |\nabla \rho|$ as the fundamental variable

$$G_{\mu\nu}^{\text{xc}}(\mathbf{M}) = \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\partial^2 f}{\partial \rho^2} \Omega_{\mu\nu} \Omega_{\rho\sigma} + \frac{\partial^2 f}{\partial \rho \partial \xi} \left(\Omega_{\mu\nu} \frac{\nabla \rho}{\xi} \nabla \Omega_{\rho\sigma} + \right. \\ \left. \Omega_{\rho\sigma} \frac{\nabla \rho}{\xi} \nabla \Omega_{\mu\nu} \right) d\mathbf{r} + \sum_{\rho\sigma} M_{\sigma\rho} \int \frac{\partial^2 f}{\partial \xi^2} \left(\frac{\nabla \rho}{\xi} \nabla \Omega_{\mu\nu} \right) \left(\frac{\nabla \rho}{\xi} \nabla \Omega_{\rho\sigma} \right) d\mathbf{r} \quad (124)$$

Differentiation of the G^{xc} matrix elements eq 124 with respect to a magnetic field (at zero magnetic field strength) gives

$$\frac{\partial G_{\mu\nu}^{\text{xc}}(\mathbf{M})}{\partial \mathbf{B}} \Big|_{\mathbf{B}=0} = \sum_{\rho\sigma} M_{\sigma\rho} \frac{i}{2} \int (\mathbf{R}_{MN} \times \mathbf{r}) \left(\frac{\partial^2 f}{\partial \rho^2} \Omega_{\mu\nu} \Omega_{\rho\sigma} + \right. \\ \frac{\partial^2 f}{\partial \rho \partial \xi} \Omega_{\mu\nu} \left(\frac{\nabla \rho}{\xi} \nabla \Omega_{\rho\sigma} \right) + \frac{\partial^2 f}{\partial \rho \partial \xi} \left(\frac{\nabla \rho}{\xi} \nabla \Omega_{\mu\nu} \right) \Omega_{\rho\sigma} + \\ \left. \frac{\partial^2 f}{\partial \xi^2} \left(\frac{\nabla \rho}{\xi} \nabla \Omega_{\mu\nu} \right) \left(\frac{\nabla \rho}{\xi} \nabla \Omega_{\rho\sigma} \right) \right) d\mathbf{r} + \sum_{\rho\sigma} M_{\sigma\rho} \frac{i}{2} \times \\ \int (\mathbf{R}_{RS} \times \mathbf{r}) \left(\frac{\partial^2 f}{\partial \rho^2} \Omega_{\mu\nu} \Omega_{\rho\sigma} + \frac{\partial^2 f}{\partial \rho \partial \xi} \Omega_{\rho\sigma} \left(\frac{\nabla \rho}{\xi} \nabla \Omega_{\mu\nu} \right) + \right. \\ \left. \frac{\partial^2 f}{\partial \rho \partial \xi} \left(\frac{\nabla \rho}{\xi} \nabla \Omega_{\rho\sigma} \right) \Omega_{\mu\nu} + \frac{\partial^2 f}{\partial \xi^2} \left(\frac{\nabla \rho}{\xi} \nabla \Omega_{\mu\nu} \right) \left(\frac{\nabla \rho}{\xi} \nabla \Omega_{\rho\sigma} \right) \right) d\mathbf{r} + \\ \sum_{\rho\sigma} M_{\sigma\rho} \frac{i}{2} \int (\mathbf{R}_{MN} \times \nabla \rho) \left(\frac{\partial^2 f}{\partial \rho \partial \xi} \frac{1}{\xi} \Omega_{\mu\nu} \Omega_{\rho\sigma} + \right. \\ \left. \frac{\partial^2 f}{\partial \xi \partial \xi} \frac{1}{\xi^2} \Omega_{\mu\nu} (\nabla \rho \nabla \Omega_{\rho\sigma}) \right) d\mathbf{r} + \\ \sum_{\rho\sigma} M_{\sigma\rho} \frac{i}{2} \int (\mathbf{R}_{RS} \times \nabla \rho) \left(\frac{\partial^2 f}{\partial \rho \partial \xi} \frac{1}{\xi} \Omega_{\mu\nu} \Omega_{\rho\sigma} + \right. \\ \left. \frac{\partial^2 f}{\partial \xi^2} \frac{1}{\xi^2} \Omega_{\rho\sigma} (\nabla \rho \nabla \Omega_{\mu\nu}) \right) d\mathbf{r} + G_{\mu\nu}^{\text{xc}} \left(\frac{\partial \mathbf{M}}{\partial \mathbf{B}} \right) \quad (125)$$

where we have used eq 96 and 101 and that all contributions from differentiation of the density itself vanish due to eq 93 and 94. The last term in eq 125 can be straightforwardly derived and is therefore omitted.

References

- (1) Faraday, M. *Philos. Mag.* **1846**, 28, 294.
- (2) Faraday, M. *Philos. Trans. R. Soc. London* **1846**, 136, 1.
- (3) Barron, L. D. *Molecular Light Scattering and Optical Activity*; 2nd ed. revised and enlarged; Cambridge University Press: Cambridge, England, 2004 pp 292.
- (4) Caldwell, D. J.; Eyring, H. The Faraday Effect. In *The Theory of Optical Activity*; Wiley-Interscience: New York, 1971; pp 167–193.
- (5) Buckingham, A. D.; Stephens, P. J. *Adv. Res. Chem. Phys.* **1966**, 17, 399.
- (6) Schatz, P. N.; McCaffery, A. J. *Q. Rev.* **1969**, 23, 552.
- (7) Sutherland, J. C. The magnetic optical activity of hemoproteins. In *The Porphyrins*; Dolphin, D., Ed.; Academic Press: New York, 1978; Vol. 1, p 56.
- (8) Dawson, J. H.; Dooley, D. M. Magnetic circular dichroism spectroscopy of iron porphyrins and heme proteins. In *Iron Porphyrins*; Lever, A.; Gray, H. P., Eds.; VCH: New York, 1989; Vol 3.
- (9) Gorski, A.; Vogel, E.; Sessler, J. L.; Waluk, J. *J. Phys. Chem. A* **2002**, 106, 8139.

- (10) Kobayashi, N.; Nakai, K. *Chem. Commun.* **2007**, 4077–4092.
- (11) Pearce, L. L.; Bominaar, E. L.; Peterson, J. *Biochem. Biophys. Res. Commun.* **2002**, 297, 220.
- (12) Verdet, E. M. *Ann. Chim. (3rd Ser.)* **1854**, 41, 370.
- (13) Kula, M.; Cappelli, C.; Coriani, S.; Rizzo, A. *ChemPhys-Chem* **2008**, 9, 462.
- (14) Botek, E.; Champagne, B.; Verbiest, T.; Gangopadhyay, P.; Persoons, A. *ChemPhysChem* **2006**, 7, 1654.
- (15) Coriani, S.; Hättig, C.; Jørgensen, P.; Halkier, A.; Rizzo, A. *Chem. Phys. Lett.* **1997**, 281, 445 Erratum, **1998**, 293, 324.
- (16) Coriani, S.; Jørgensen, P.; Christiansen, O.; Gauss, J. *Chem. Phys. Lett.* **2000**, 330, 463.
- (17) Coriani, S.; Hättig, C.; Jørgensen, P.; Helgaker, T. *J. Chem. Phys.* **2000**, 113, 3561.
- (18) Banerjee, A.; Autschbach, J.; Ziegler, T. *Int. J. Quantum Chem.* **2005**, 101, 572.
- (19) Krykunov, M.; Banerjee, A.; Ziegler, T.; Autschbach, J. *J. Chem. Phys.* **2005**, 122, 074105.
- (20) Michl, J.; Thulstrup, E. W. *Spectroscopy with Polarized Light*; VCH Publishers, Inc.: New York, 1986.
- (21) Thulstrup, E. W. *Aspects of the Linear Magnetic Circular Dichroism of Planar Organic Molecules*; Springer-Verlag: Berlin, 1980.
- (22) Mason, W. R. *A Practical Guide to Magnetic Circular Dichroism Spectroscopy*; John Wiley and Sons: New York, 2007.
- (23) Piepho, S. B.; Schatz, P. N. *Group Theory in Spectroscopy: With Applications to Magnetic Circular Dichroism*, John Wiley and Sons: New York, 1983.
- (24) Solomon, E.; Pavel, E.; Loeb, K.; Campochiaro, C. *Coord. Chem. Rev.* **1995**, 144, 369.
- (25) Kirk, M.; Peariso, K. *Curr. Opin. Chem. Biol.* **2003**, 7, 220.
- (26) Cheeseman, M.; Greenwood, C.; Thomson, T. J. *Advances in Inorganic Chemistry*; Academic Press: San Diego, CA, 1991; Vol. 36, pp 201–255.
- (27) Serber, R. *Phys. Rev.* **1932**, 41.
- (28) Stephens, P. J. *Chem. Phys. Lett.* **1968**, 2, 241.
- (29) Stephens, P. J. *J. Chem. Phys.* **1970**, 52, 3489.
- (30) Stephens, P. J. *Annu. Rev. Phys. Chem.* **1974**, 25, 201.
- (31) Stephens, P. J. *Adv. Chem. Phys.* **1976**, 35, 197.
- (32) Solheim, H.; Ruud, K.; Coriani, S.; Norman, P. J. *Phys. Chem. A* **2008**, 112, 9615.
- (33) Seamans, L.; Moscovitz, A. *J. Chem. Phys.* **1972**, 56, 1099.
- (34) Coriani, S.; Jørgensen, P.; Rizzo, A.; Ruud, K.; Olsen, J. *Chem. Phys. Lett.* **1999**, 300, 61.
- (35) Honda, Y.; Hada, M.; Ehara, M.; Nakatsuji, H.; Michl, J. *J. Chem. Phys.* **2005**, 123, 164113.
- (36) Kjærgaard, T.; Jansík, B.; Jørgensen, P.; Coriani, S.; Michl, J. *J. Phys. Chem. A* **2007**, 111, 11278.
- (37) Solheim, H.; Frediani, L.; Ruud, K.; Coriani, S. *Theor. Chem. Acc.* **2008**, 119, 231.
- (38) Solheim, H.; Ruud, K.; Coriani, S.; Norman, P. J. *Chem. Phys.* **2008**, 128, 094103.
- (39) Seth, M.; Ziegler, T.; Banerjee, A.; Autschbach, J.; van Gisbergen, S.; Baerends, E. *J. Chem. Phys.* **2004**, 120, 10942.
- (40) Seth, M.; Ziegler, T.; Autschbach, J. *J. Chem. Phys.* **2005**, 122, 094112.
- (41) Peralta, G. A.; Seth, M.; Ziegler, T. *J. Chem. Theory Comput.* **2007**, 3, 434.
- (42) Seth, M.; Krykunov, M.; Ziegler, T.; Autschbach, J.; Banerjee, A. *J. Chem. Phys.* **2008**, 128, 144105.
- (43) Peralta, G. A.; Seth, M.; Ziegler, T. *Inorg. Chem.* **2007**, 46, 9111.
- (44) Krykunov, M.; Seth, M.; Ziegler, T.; Autschbach, J. *J. Chem. Phys.* **2007**, 127, 244102.
- (45) Ganyushin, D.; Neese, F. *J. Chem. Phys.* **2008**, 128, 114117.
- (46) Seth, M.; Ziegler, T. *Inorg. Chem.* **2009**, 48, 1793.
- (47) Jørgensen, P.; Oddershede, J.; Beebe, N. H. F. *J. Chem. Phys.* **1978**, 68, 2527.
- (48) Jaszuński, M.; Jørgensen, P.; Rizzo, A.; Ruud, K.; Helgaker, T. *Chem. Phys. Lett.* **1994**, 222, 263.
- (49) Parkinson, W. A.; Sauer, S. P. A.; Oddershede, J.; Bishop, D. M. *J. Chem. Phys.* **1993**, 98, 487.
- (50) Jaszuński, M.; Jørgensen, P.; Rizzo, A. *Theor. Chim. Acta* **1995**, 90, 291.
- (51) Coriani, S.; Hättig, C.; Rizzo, A. *J. Chem. Phys.* **1999**, 111, 7828.
- (52) London, F. *J. Phys. Radium* **1937**, 8, 397.
- (53) Helgaker, T.; Jørgensen, P. *J. Chem. Phys.* **1991**, 95, 2595.
- (54) Bak, K. L.; Hansen, A. E.; Ruud, K.; Helgaker, T.; Olsen, J.; Jørgensen, P. *Theor. Chim. Acta* **1995**, 90, 441.
- (55) Miles, D. W.; Eyring, H. *Proc. Natl. Acad. Sci. U.S.A.* **1973**, 70, 3754.
- (56) Meier, A. R.; Wagnière, G. H. *Chem. Phys.* **1987**, 113, 287.
- (57) Shieh, D. J.; Lin, S. H.; Eyring, H. *J. Phys. Chem.* **1972**, 76, 1844.
- (58) Goldstein, E.; Vijaya, S.; Segal, G. A. *J. Am. Chem. Soc.* **1980**, 102, 6198.
- (59) Marconi, G. *Chem. Phys. Lett.* **1988**, 146, 259.
- (60) Michl, J. *Tetrahedron* **1974**, 30, 4215.
- (61) Fleischhauer, J.; Michl, J. *J. Phys. Chem. A* **2000**, 104, 7776.
- (62) Caldwell, D.; Eyring, H. *J. Chem. Phys.* **1973**, 58, 1149.
- (63) Kaito, A.; Tajiri, A.; Hatano, M. *J. Am. Chem. Soc.* **1976**, 96, 384.
- (64) Sprinkel, F. M.; Shillady, D. D.; Strickland, R. W. *J. Am. Chem. Soc.* **1975**, 97, 6653.
- (65) Seth, M.; Krykunov, M.; Ziegler, T.; Autschbach, J. *J. Chem. Phys.* **2008**, 128, 234102.
- (66) Seth, M.; Ziegler, T. *J. Chem. Phys.* **2007**, 127, 134108.
- (67) Seamans, L.; Linderberg, J. *Mol. Phys.* **1972**, 24, 1393.
- (68) Gross, E. K. U.; Dobson, J. F.; Petersilka, M. Density functional theory of time-dependent phenomena. In *Topics in Current Chemistry*; Springer: Berlin Heidelberg, 1996; Vol. 181, pp 82.
- (69) van Leeuwen, R. *Int. J. Mod. Phys. B* **2001**, 15, 1969.
- (70) Koch, W.; Holthausen, M. C. A. *Chemist's Guide to Density Functional Theory*, 2nd ed.; Wiley-VCH: Weinheim, Germany, 2001.
- (71) Sauek, P.; Høst, S.; Thøgersen, L.; Jørgensen, P.; Manninen, P.; Olsen, J.; Jansík, B.; Reine, S.; Pawłowski, F.; Tellgren,

- E.; Helgaker, T.; Coriani, S. *J. Chem. Phys.* **2007**, *126*, 114110.
- (72) Rubensson, E. H.; Salek, P. *J. Comput. Chem.* **2005**, *26*, 1628.
- (73) Jansík, B.; Høst, S.; Jørgensen, P.; Olsen, J.; Helgaker, T. *J. Chem. Phys.* **2007**, *126*, 124104.
- (74) Coriani, S.; Høst, S.; Jansík, B.; Thøgersen, L.; Olsen, J.; Jørgensen, P.; Reine, S.; Pawłowski, F.; Helgaker, T.; Salek, P. *J. Chem. Phys.* **2007**, *126*, 154108.
- (75) Kjærsgaard, T.; Jørgensen, P.; Olsen, J.; Coriani, S.; Helgaker, T. *J. Chem. Phys.* **2008**, *129*, 054106.
- (76) Larsen, H.; Jørgensen, P.; Olsen, J.; Helgaker, T. *J. Chem. Phys.* **2000**, *113*, 8908.
- (77) Coriani, S.; Kjærsgaard, T.; Jørgensen, P.; Ruud, K.; Berger, R. Unpublished work.
- (78) Furche, F.; Ahlrichs, R. *J. Chem. Phys.* **2002**, *117*, 7433.
- (79) Rappoport, D.; Furche, F. *J. Chem. Phys.* **2007**, *126*, 201104.
- (80) Thorvaldsen, A.; Ruud, K.; Kristensen, K.; Jørgensen, P.; Coriani, S. *J. Chem. Phys.* **2008**, *129*, 214108.
- (81) Shcherbin, D.; Thorvaldsen, A. J.; Ruud, K.; Coriani, S.; Rizzo, A. *Phys. Chem. Chem. Phys.* **2009**, *11*, 816.
- (82) Buckingham, A. D.; Jamieson, M. J. *Mol. Phys.* **1971**, *22*, 117.
- (83) Olsen, J.; Jørgensen, P. *J. Chem. Phys.* **1985**, *82*, 3235.
- (84) Olsen, J.; Jørgensen, P. Time-dependent response theory with application to self-consistent field and multiconfigurational self-consistent field wave functions. In *Modern Electronic Structure Theory*, Part II; Yarkony, D. R., Ed.; World Scientific: Singapore, 1995; Vol 2, pp 857.
- (85) Helgaker, T.; Wilson, P. J.; Amos, R. D.; Handy, N. C. *J. Chem. Phys.* **2000**, *113*, 2983.
- (86) Helgaker, T.; Jørgensen, P.; Olsen, J. *Molecular Electronic—Structure Theory*; Wiley: Chichester, U.K., 2000; pp 468.
- (87) Larsen, H.; Helgaker, T.; Jørgensen, P.; Olsen, J. *J. Chem. Phys.* **2001**, *115*, 10344.
- (88) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007.
- (89) Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1993**, *98*, 1358.
- (90) Woon, D. E.; Dunning, T. H., Jr. *J. Chem. Phys.* **1994**, *100*, 2975.
- (91) Ingersoll, L. R.; Liebenberg, D. H. *J. Opt. Soc. Am.* **1954**, *44*, 566.
- (92) Ingersoll, L. R.; Liebenberg, D. H. *J. Opt. Soc. Am.* **1956**, *46*, 538.
- (93) Vosko, S.; L. Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200.
- (94) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- (95) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51.
- (96) Peach, M. J. G.; Helgaker, T.; Salek, P.; Keal, T. W.; Lutnæs, O. B.; Tozer, D. J.; Handy, N. C. *Phys. Chem. Chem. Phys.* **2006**, *8*, 558.
- (97) Bishop, D. M.; Cybulski, S. *J. Chem. Phys.* **1990**, *93*, 590.
- (98) Mort, B.; Autschbach, J. *J. Phys. Chem. A* **2007**, *111*, 5563–5571.
- (99) Castellani, A.; Michl, J. *J. Am. Chem. Soc.* **1978**, *100*, 6824.
- (100) Waluk, J.; Klein, H.; Ashe, A. J.; Michl, J. *Organometallics* **1989**, *8*, 2804.
- (101) Peach, M. J. G.; Benfield, P.; Helgaker, T.; Tozer, D. J. *J. Chem. Phys.* **2008**, *128*, 044118.
- (102) Parr, R. G.; Yang, W. *Density-Functional Theory of Atoms and Molecules*; Oxford Science Publications: Oxford, U.K., 1989.
- (103) Salek, P.; Hesselmann, A. *J. Comput. Chem.* **2007**, *28*, 2569.

CT9001625

Theoretical Investigation of the Reaction of Imidogen with Fulminic Acid

Ya-nan Xin, Min Zhao,* and Zuo-sheng Li

*College of Chemistry and Chemical Engineering, Bohai University,
Jinzhou, Liaoning, 121000, People's Republic of China*

Wei Xiong, Xinli Song, Hua Hou, and Baoshan Wang*

*College of Chemistry and Molecular Sciences, Wuhan University, Wuhan 430072,
People's Republic of China*

Received May 7, 2009

Abstract: The mechanism of the reaction of imidogen (NH) with fulminic acid (HCNO) has been investigated theoretically using the multiconfigurational self-consistent-field theory (MCSCF), multireference Rayleigh–Schrodinger perturbation theory (RSPT2), and coupled cluster theory (CC) along with the complete basis set extrapolations (CBS). The calculations show that the NH + HCNO reaction takes place via an N → C addition mechanism predominantly by surmounting a small barrier (ca. ~3 kcal/mol). The adduct is HC(NH)NO in the triplet state with an exothermicity of more than 60 kcal/mol. The subsequent C–N cleavage, which is nearly barrierless, leads to HCNH and NO as the final products. This represents the most energetically favorable product channel of the title reaction. The channels leading to HCN, HNC, HNO, or HON via O- or H-migration mechanisms involve higher barriers and thus are negligible. The singlet–triplet crossing has been investigated as well for the HCNH + NO product channel by locating the conical interactions. Using transition state theory, the rate constants were predicted as a function of temperatures. It is suggested that the NH + HCNO reaction might be an alternative source for the NO regeneration under the combustion conditions. This calculation is useful to simulate experimental investigations of the NH + HCNO reaction.

I. Introduction

In contrast to its isomer isocyanic acid (HNCO), fulminic acid (HCNO) is less stable but can be synthesized and stored for brief periods in the gas phase.^{1,2} The structure^{1–4} and reactivity^{5–11} of this molecule have attracted considerable attention both experimentally and theoretically. Under combustion conditions, HCNO is formed primarily from the acetylene oxidation followed by the HCCO + NO reaction. It plays an important role in the so-called NO-reburning process for the reduction of NO_x pollutants from the fossil-fuel emission.¹¹ Very recently, it has been discovered that HCNO also exists in dark clouds with an abundance of 1–5 × 10⁻¹⁰ with respect to molecular hydrogen. The CH₂ +

NO → HCNO + H reaction was suggested to be a key path for the formation of fulminic acid.¹²

The reactions of HCNO with a few highly reactive radicals, for example, H, O, OH, CN, NCO, etc., have been studied extensively.^{5–11} Most of these reactions are significantly exothermic and relatively fast, with rate constants on the order of 10⁻¹¹ cm³ molecule⁻¹ s⁻¹. Theoretical calculations of these reactions have been reported using various quantum chemistry methods, for example, density functional theory and the coupled cluster theory.^{10,11} The predicted reaction mechanisms are in agreement with the experimental observations.

NH(X³Σ⁻) is an important radical in both combustion and atmospheric chemistry.^{13–15} Although its reactivity is lower than the radicals (e.g., OH) as mentioned above, the NH

* Corresponding authors e-mail: baoshan@whu.edu.cn.

radical can play a critical role in the NO-reburning process. In this work, the reaction of NH with HCNO is investigated theoretically using the high-level ab initio quantum chemistry methods. The theoretical data are invaluable for extending our understanding of the roles of NH and HCNO in both combustion and atmosphere. To our knowledge, the NH + HCNO reaction has not been studied either experimentally or theoretically. The present calculations might shed new light on the importance of this reaction in the gas phase.

II. Computational Methods

As a preliminary effort, the geometries of the reactants, products, intermediates, and transition states were fully optimized using the conventional B3LYP¹⁶ and MP2¹⁷ methods with the standard 6-311+G(d,p)¹⁸ basis set using the Gaussian03 programs.¹⁹ However, it was found that the single-reference wave functions are subject to severe spin contaminant. For instance, the spin expectation value $\langle S^2 \rangle$ for the association transition state TS1 is as high as 2.352 at the UMP2/6-311+G(d,p) level, whereas the correct value for the triplet is 2.0. Even for the HC(NH)NO adduct, the value of $\langle S^2 \rangle$ is 2.15. Moreover, the problem of spin contaminant cannot be conquered by using either higher order perturbation theory (e.g., MP4) or configuration interaction theory (e.g., CISD). These preliminary calculations clearly indicate that the multireference and more extensive electronic correlation methods are required to explore the energetic reaction routes of the NH + HCNO reaction.

Two high-level methods have been employed in the geometrical optimizations. The first method is the multiconfigurational self-consistent-field method (MCSCF).^{20,21} The full valence active space includes 22 electrons distributed in 18 orbitals. For the sake of our computational resources, we chose a smaller active space, namely, 14 electrons in 12 orbitals, consisting of 141 760 configuration state functions (CSFs). It has been carefully checked that this active space is capable of including all the important CSFs for the stationary points (for example, see Figure S1 in the Supporting Information) and the MCSCF wave function converges to the correct ground state. The “supermolecule” approximation was used for the two fragments of the reactants and the products with a separation of 100 Å. The second method is the coupled cluster theory with single, double, and noniterative triple excitations, namely, CCSD(T).²² For the triplet species of concern, the restricted open-shell wavefunction was used for both CCSD(T) and HF references, namely, RCCSD(T,full)-ROHF, where the term “full” means that all electrons are included in the electronic correlation calculations. The diagnostic T_1 values for the species have been monitored to be in the range 0–0.04, indicating the good quality of the RCCSD(T,full)-ROHF wave functions.²³ The 6-311+G(d,p) basis set was used for both methods. The convergency criterion requires the maximum component of the gradient to be less than 1.0 × 10⁻⁶ au in all optimization calculations.^{24,25}

Harmonic vibrational frequencies for all stationary points were calculated at the same level of theory as used in the optimization [e.g., MCSCF and RCCSD(T,full)] by numerically approximating Hessian.^{26,27} The minimum has all real

frequencies. The transition state only involves one imaginary frequency representing the motion along the reaction coordinate. Moreover, a few intrinsic reaction coordinate (IRC)²⁸ calculations were carried out in order to confirm that the connections of the transition states with the designated reactants and products are valid.

On the basis of the MCSCF(14e,12o)/6-311+G(d,p)-optimized geometries, the single-point energy calculations were carried out using the multireference Rayleigh–Schrodinger perturbation theory.^{29,30} More accurate multireference configuration interaction (MRCI) calculation is not affordable at present for the large (14e,12o) active space. The internally contracted version of RS2C has been chosen in this work. Moreover, a level shift of 0.2 was used in order to avoid intruder state problem in excited state calculations as suggested by Roos and Andersson.³¹ Two correlation-consistent basis sets, namely, aug-cc-pVDZ and aug-cc-pVTZ,³² were used for RS2C. The total numbers of contracted configurations are 28 709 896 and 68 432 064, respectively. The individual energies were extrapolated to the complete basis set limit (CBS) using the following expression³³

$$E_X = E_\infty + cX^{-3} \quad (X = 2, 3) \quad (1)$$

where X represents the cardinal number X of the aug-cc-pVXZ basis sets.

On the basis of the RCCSD(T,full)/6-311+G(d,p)-optimized geometries, further RCCSD(T,full) single-point calculations were done using the aug-cc-pVDZ and aug-cc-pVTZ basis sets. Moreover, an additional ROHF single-point energy was obtained using the aug-cc-pVQZ basis set. Then, the full coupled-cluster energy (FCC) was calculated using the following expressions:^{34,35}

$$E(\text{FCC}) = E_\infty[\text{CCSD(T)}] + 1/5\Delta E_{\text{aug-cc-pVTZ}}^T \quad (2)$$

$$E_\infty[\text{CCSD(T)}] = E_\infty(\text{HF}) + \Delta E_\infty[\text{CCSD(T)}] \quad (3)$$

$$E_X(\text{HF}) = E_\infty(\text{HF}) + a \exp(-bX) \quad (4)$$

$$\Delta E_X[\text{CCSD(T)}] = \Delta E_\infty[\text{CCSD(T)}] + cX^{-3} \quad (5)$$

In addition, the extrapolation scheme developed by Truhlar was used for the purpose of comparison and cross-validation of the CBS extrapolations, viz.³⁶

$$E_\infty[\text{CCSD(T)}] = \frac{3^\alpha E_{\text{HF},3} - 2^\alpha E_{\text{HF},2}}{3^\alpha - 2^\alpha} + \frac{3^\beta E_{\text{corr},3} - 2^\beta E_{\text{corr},2}}{3^\beta - 2^\beta} \quad (6)$$

where $\alpha = 3.4$ and $\beta = 2.4$. All of the above ab initio calculations were carried out using the Molpro2006.1 program.³⁷

III. Results and Discussion

The optimized geometries of the reactants, products, intermediates (INT), and transition states (TS) at the MCSCF(14e,12o) and RCCSD(T,full) levels of theory are shown comparatively in Figure 1. The key energetic reaction routes are shown

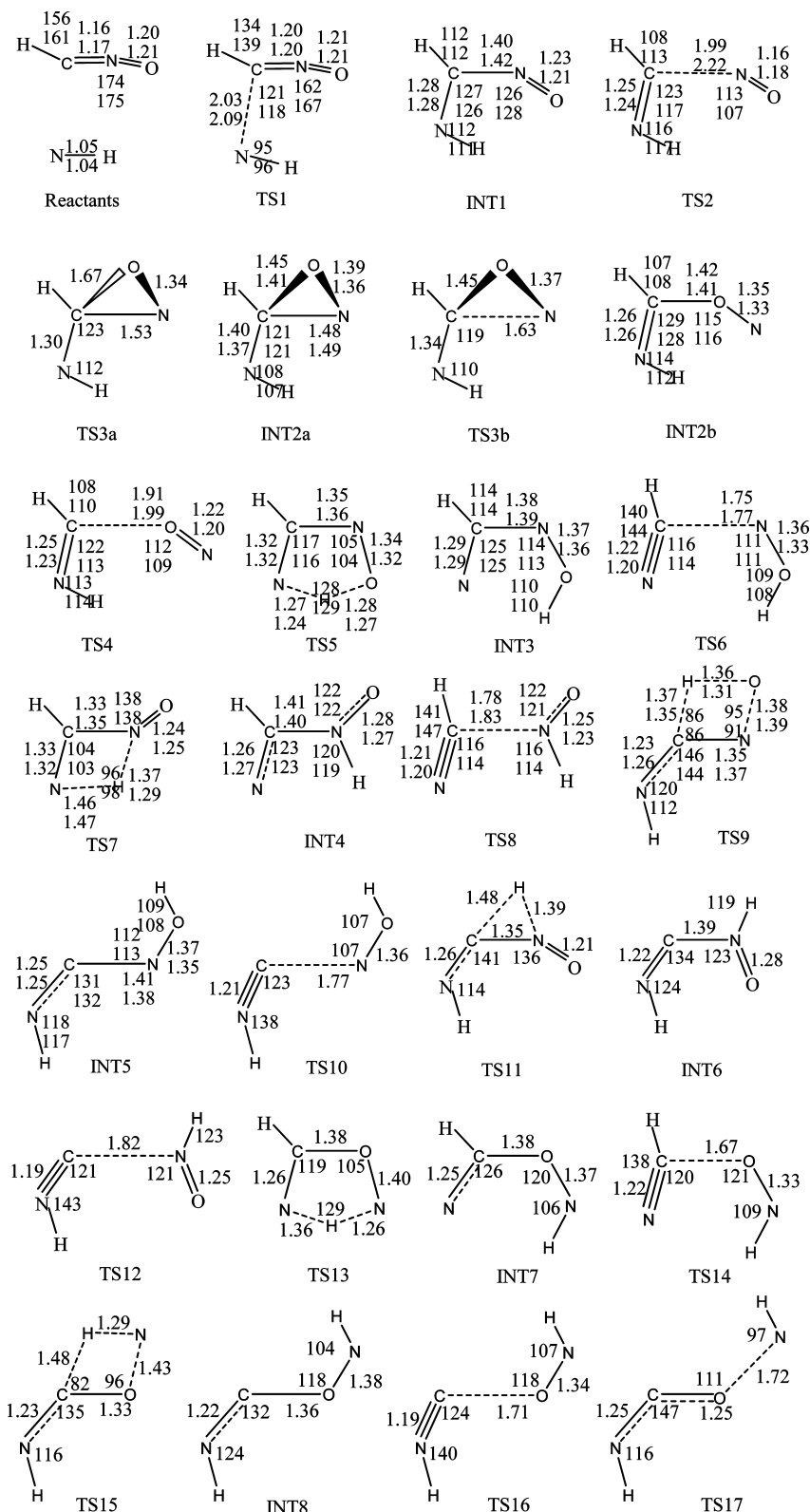


Figure 1. Geometries of the transition states (TS) and intermediates (INT) involved in the NH + HCNO reaction. Bond distances are in Ångstroms, and bond angles are in degrees. Upper entries: the MCSCF(14e,12o)/6-311+G(d,p)-optimized parameters. Lower entries: the RCCSD(T,full)/6-311+G(d,p)-optimized parameters.

schematically in Figure 2. The zero-point energies (ZPE) and relative energies for all stationary points are summarized in Table 1. More data including vibrational frequencies (Table S1 in the Supporting Information) and full reaction mechanisms calculated at the G4//MP2(full)/6-311+G(d,p) level

(Figure S2 in the Supporting Information) are deposited as Supporting Information.

The effect of the sizes of the basis sets on the geometrical parameters was carefully checked as well using 6-311+G(2d,2p), 6-311+G(3df,3pd), aug-cc-pVDZ, and aug-cc-pVTZ basis

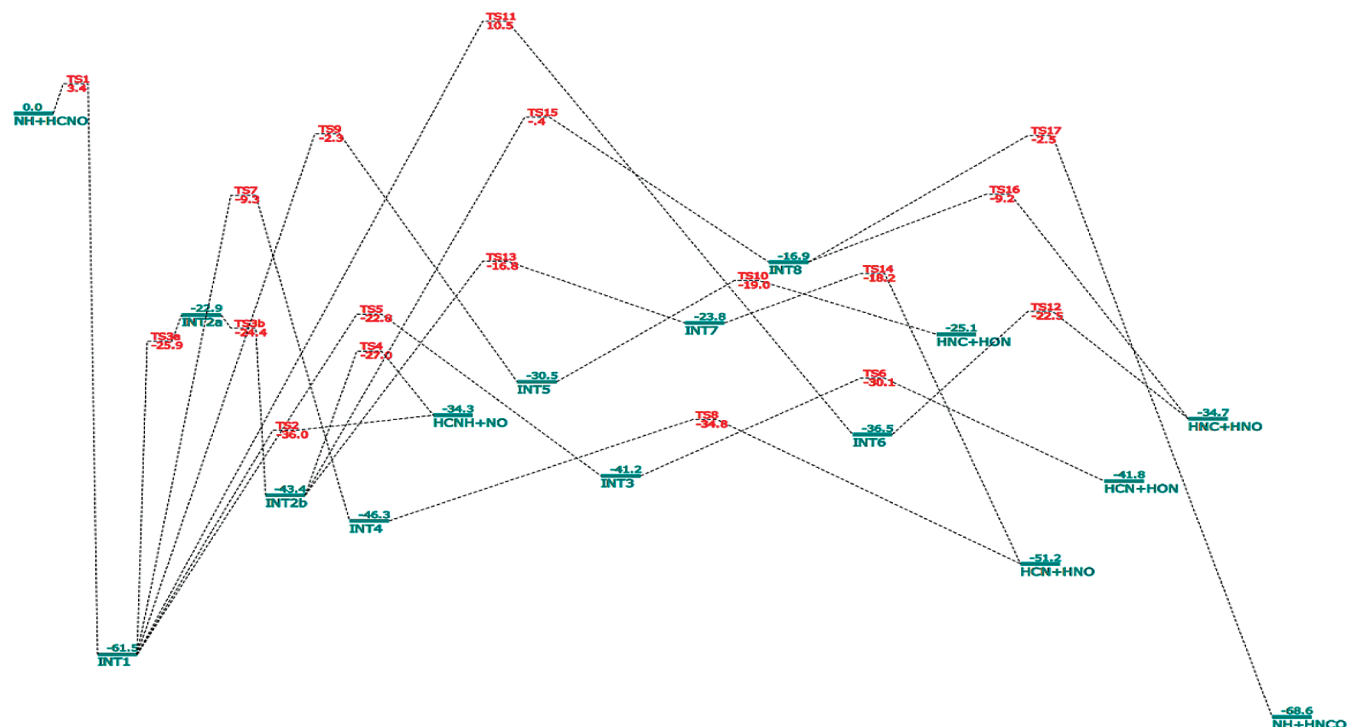


Figure 2. Energetic reaction routes of the NH + HCNO reaction. The energies calculated at the RS2C/CBS//MCSCF(14e,12o)/6-311+G(d,p) level are shown in units of kcal/mol.

sets for the association transition state (TS1) and intermediate (INT1). These data are shown in the Supporting Information (Tables S2 and S3). It is evident that the basis set employed in this work, i.e., 6-311+G(d,p), is adequate for the geometrical optimizations.

The geometry of fulminic acid is still uncertain. It was suggested by the high-level ab initio calculations that HCNO has a bent geometry, but the collinear geometry is only a few wavenumbers higher in energy.^{2–4} Our results obtained at both MCSCF and RCCSD(T) levels are in good agreement with previous calculations. Moreover, the RCCSD(T) vibrational frequencies are in good agreement with the experimental infrared measurements (see Table S1 in the Supporting Information).¹ In fact, HCNO can be seen as a resonant structure between H–C≡N–O and H–C=N=O.

There are four possible reaction sites while NH is approaching toward the H, C, N, and O atoms of HCNO. As shown in Figure S2 in the Supporting Information, the H abstraction and the N- and O-addition pathways are endothermic and involve significant barriers. Therefore, they are negligible. In the following section, only the C-addition reaction mechanism will be discussed in detail. Note that most of the stationary points are shown to be planar structures and possess $^3A''$ electronic states in the C_s symmetry, in accordance with the adiabatic correlation asymptote of $NH(X^3\Sigma^-) + HCNO(X^1A')$.

1. N–C Addition Path. The association reaction occurs between the N atom of NH and the C atom of HCNO. The singly occupied p orbital of NH is approaching the Π orbitals of HCNO to form the CN bond (see Figure S1 in the Supporting Information). The adduct is denoted as INT1. In view of the spin densities, the Π orbital of HCNO is broken totally because the two radical centers move to N and O

atoms, respectively. The newly formed CN bond exhibits double-bonding character, and the other C–N bond is a single bond. The addition step is highly exothermic by more than 60 kcal/mol.

The transition state for association is shown as TS1 in Figure 1. The Π orbital of HCNO is only slightly affected by the approaching p orbital of the N atom in NH. Evidently, it is an early barrier. The approaching C...N distance is about 0.8 Å longer than that in INT1. Note that TS1 is the only true transition state for the association even though the adduct INT1 has an isomer with the *cis*-HCNO conformation. The geometrical parameters obtained at the MCSCF level are in reasonable agreement with those at RCCSD(T) level, although RCCSD(T) tends to give an earlier transition state.

It is extremely difficult to calculate the barrier height for such a loose barrier. At the RS2C/CBS//MCSCF(14e,12o)/6-311+G(d,p) level, the classical barrier height (e.g., without ZPE) becomes negative, namely, –3.4 kcal/mol. However, at the FCC/CBS//RCCSD(T,full)/6-311+G(d,p) level, the classical barrier is 1.4 kcal/mol. If the RCCSD(T,full)/6-311+G(d,p)-optimized geometrical parameters were used, the calculated RS2C/CBS barrier height increases to –1.5 kcal/mol. The barrier height seems to be very sensitive to the geometry of the transition state.

In order to clarify what happens in these calculations, we performed an IRC calculation of TS1 at the MCSCF(6e,6o)/6-311+G(d,p) level of theory. For every point along the obtained minimum energy reaction path, we calculated the high-level energies using RS2C, CASMP2,³⁸ MRCI, and MRCI+Q methods^{39,40} based on the MCSCF(6,6) configurations. The results are shown in Figure 3. It is obvious that the position of the barrier depends on the levels of theory.

Table 1. Relative Energies (in kcal/mol) for the Important Species Involved in the NH + HCNO Reaction^a

species	$\Delta ZPE_{\text{MCSCF}}$	ΔE_{MCSCF}	ΔE_{RS2C}	$\Delta ZPE_{\text{RCCSD(T)}}$	T_1^f	$\Delta E_{\text{RCCSD(T)}}$	ΔE_{FCC}	$\Delta E_{\text{FCC-Truhlar}}$
TS1	2.5	2.8	-3.4 -1.5 ^b 0.4 ^c 1.3 ^d 2.0 ^e	2.1	0.024	4.1	1.4	1.4
INT1	7.7	-43.0	-68.8	7.0	0.025	-63.4	-68.0	-67.5
TS2	4.7	-15.5	-40.3	3.2	0.023	-39.0	-39.4	-38.6
HCNH + NO	2.9	-44.5	-36.9	2.4		-41.4	-38.4	-37.2
TS3a	5.1	-10.0	-31.0					
INT2a	6.0	-12.2	-28.6	5.6	0.026	-26.8	-29.6	-29.7
TS3b	5.3	-8.8	-29.4					
INT2b	7.0	-39.1	-50.1	6.4	0.019	-50.1	-50.9	-50.6
TS4	4.0	-14.1	-30.7	3.3	0.034	-28.8	-30.7	-29.8
TS5	2.9	4.6	-25.3	2.9	0.027	-18.8	-24.9	-24.5
INT3	5.9	-26.8	-46.7	5.7	0.025	-42.9	-45.4	-44.7
TS6	4.5	-18.2	-34.3	3.7	0.020	-29.9	-32.4	-32.0
HCN + NOH	2.6	-48.8	-44.1	2.2		-46.8	-42.0	-41.3
TS7	3.0	18.7	-12.0	2.9	0.034	-1.8	-7.9	-7.5
INT4	6.4	-26.2	-52.3	6.2	0.025	-49.1	-52.2	-51.8
TS8	3.5	-19.9	-37.9	2.9	0.039	29.8	-32.4	-32.1
HCN + HNO	2.3	-54.4	-53.2	1.9		-51.7	-50.6	-50.1
TS9	2.5	30.4	-4.5	2.2	0.035	4.4	-0.4	-0.3
INT5	6.4	-10.4	-36.5	5.5	0.027	-31.6	-35.5	-35.0
TS10	4.2	-0.8	-22.8		0.027		-17.8 ^g	-17.3 ^g
HNC + HON	2.4	-26.8	-27.2	1.9		-31.5	-26.4	-25.5
TS11	1.8	62.8	9.0					
INT6	6.8	-18.6	-43.0					
TS12	4.2	-0.4	-26.4					
HNC + HNO	2.1	-33.5	-36.4					
TS13	3.1	5.7	-19.6	3.0	0.031	-14.9	-17.6	-17.5
INT7	6.1	-19.2	-29.6	5.1	0.021	-28.6	-30.9	-30.9
TS14	3.8	-7.3	-21.6	3.3	0.039	-16.9	-18.8	-18.4
TS15	2.6	32.4	-2.7					
INT8	5.8	-4.9	-22.3					
TS16	3.7	15.5	-12.6					
TS17	2.9	24.1	-5.0					
NH + HNCO	1.8	-69.6	-70.0					

^a All energies are calculated with respect to the NH + HCNO asymptote. $\Delta ZPE_{\text{MCSCF}}$ and ΔE_{MCSCF} : zero-point energies and electronic energies calculated at the MCSCF(14e,12o)/6-311+G(d,p) level of theory. ΔE_{RS2C} : energies calculated at the RS2C/CBS//MCSCF(14e,12o)/6-311+G(d,p) level of theory. $\Delta ZPE_{\text{RCCSD(T)}}$ and $\Delta E_{\text{RCCSD(T)}}$: zero-point energies and electronic energies calculated at the RCCSD(T,full)/6-311+G(d,p) level of theory. ΔE_{FCC} and $\Delta E_{\text{FCC-Truhlar}}$: energies calculated at the extrapolated full coupled-cluster theory using expressions E4 and E5, respectively. ^b Calculated at the RS2C/CBS//RCCSD(T,full)/6-311+G(d,p) level. ^c Calculated at the RS2C/CBS//MRCI(6e,6o)/6-311+G(d,p) level. ^d Calculated at the RS2C/CBS//MRCI(6e,6o)+Q/6-311+G(d,p) level. ^e Calculated at the RS2C/CBS//RS2C(6e,6o)/6-311+G(d,p) level. ^f T_1 diagnostic values calculated at the RCCSD(T,full)/AVTZ level. ^g The RCCSD(T,full)/6-311+G(d,p) optimization cannot converge. The energies were calculated using the MCSCF(14e,12o)/6-311+G(d,p)-optimized geometry.

At the RS2C/CBS level, the transition state occurs at $R(\text{C}\cdots\text{N}) = 2.25 \text{ \AA}$, which is significantly longer than that obtained at the MCSCF level. This explains why the RS2C-calculated barrier height at the MCSCF-optimized geometry is even negative. Both MRCI and MRCI+Q methods predict longer $R(\text{C}\cdots\text{N})$ as well. It implies that besides the nondynamical effect, the dynamical electronic correlation is also important in the determination of TS1. The noncontinuous energies at the CASMP2 level are due to the so-called “state intrude” problem. This verifies that the shift parameters in our RS2C calculations are important to obtain reasonable energies. The RS2C/CBS-calculated barrier heights are 0.4, 1.3, and 2.0 kcal/mol using the MRCI, MRCI+Q, and RS2C maxima, respectively. These values are in reasonable with those obtained at the FCC/CBS//RCCSD(T,full)/6-311+G(d,p) levels. With the above investigations, we feel confident that the barrier for the association path does exist, although it is very low. The classical barrier should be in the range 1–2 kcal/mol. After the ZPE correction, the barrier height is estimated to be 3.5 kcal/mol.

2. INT1 \rightarrow HCNH + NO Dissociation Path. As mentioned above, due to the addition of NH to the C atom of HCNO, the CN bond is significantly stretched to 1.4 \AA . Meanwhile, the association results in plenty of energy to INT1 and the C–N bond is readily broken to produce HCNH and NO.

We found a transition state TS2 for this CN bond scission process. As shown in Figure 1, TS2 is a very loose transition state. The $\text{C}\cdots\text{N}$ distance is nearly 2.0 \AA at the MCSCF level, and it is even longer if the dynamical electronic correlation has been included using the RCCSD(T) method. In fact, the energy of TS2 is somewhat lower than that of the products at both RS2C/CBS and FCC/CBS levels. Apparently, $\text{C}\cdots\text{N}$ fission would occur without any well-defined transition state. The formation of HCNH + NO is highly exothermic by 36 kcal/mol.

Note that INT1 has a conformer with the *cis*-HCNO geometry. Its energy is ~ 1 kcal/mol higher than that of INT1. Interestingly, the $\text{C}\cdots\text{N}$ fission of the *cis*-INT1 does involve

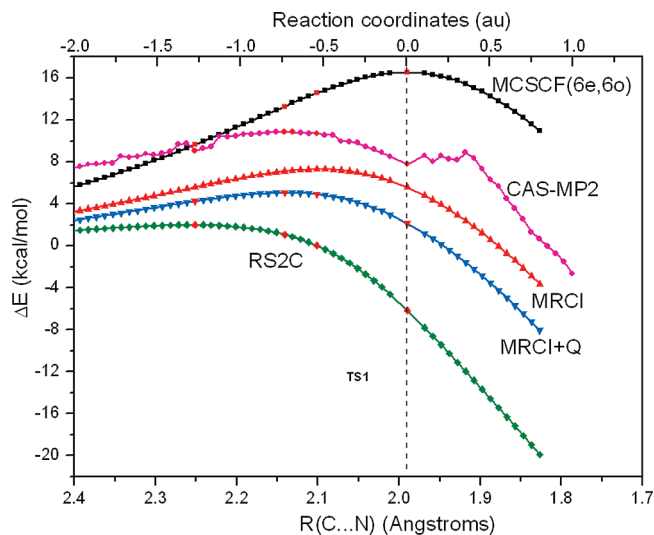


Figure 3. Minimum energy path for the association reaction of NH with HCNO. All relative energies (in kcal/mol) are calculated at the MCSCF(6e,6o)/6-311+G(d,p) level with respect to the NH + HCNO asymptote. On the basis of the geometries of each point along the MEP, the single-point energies were calculated using the CASMP2, MRCI, and RS2C methods with the aug-cc-pVDZ and aug-cc-pVTZ basis sets and then the energies were extrapolated to the complete basis set limit using eq 1. The term “Q” in the MRCI+Q method means the empirical Davidson correction to the MRCISD energy.

a well-defined barrier whose energy is about 1 kcal/mol higher than that of the products.

3. O-Rearrangement Channel Starting from INT1.

Since the ground state of NH has $3\Sigma^-$ symmetry, the NH + HCNO asymptote adiabatically correlates to the $3A''$ electronic state in the C_s point group. However, an oxygen-shift reaction route exists starting from INT1 ($3A''$) by breaking the C_s symmetry. The transition state is TS3a as shown in Figure 1. The terminal O atom is bent to the central C atom, forming a three-membered cyclic structure, INT2a. Interestingly, the classical barrier for this process is only 38 kcal/mol, which is even lower than those for the H-shift paths as will be discussed below. TS3a is the second lowest barrier in the INT1 reaction. However, the CON ring in INT2a is very unstable. The CN bond is readily broken to form a new planar intermediate, namely, INT2b. In comparison with INT1, the two singly occupied electrons in INT2b are localized on the terminal N atom. The energy of INT2b is 18 kcal/mol higher than that of INT1. The CO bond cleavage of INT2b is undergone via TS4 with the formation of HCNH + NO. TS4 appears to be a tighter barrier than TS2, although their geometrical parameters are similar. As a result, TS4 is a well-defined transition state. Its energy is about 7 kcal/mol higher than that of the products.

4. Singlet–Triplet Crossing for the HCNH + NO Channel.

The reaction mechanism for the HCNH(X^2A') + NO($X^2\Pi$) channel could be more complicated if the surface crossing was considered because this channel does correlate adiabatically with both triplet and singlet surfaces in view of the two doublet product molecules. Using the state-averaged MCSCF(14e,12o)/6-311+G(d,p) method, the con-

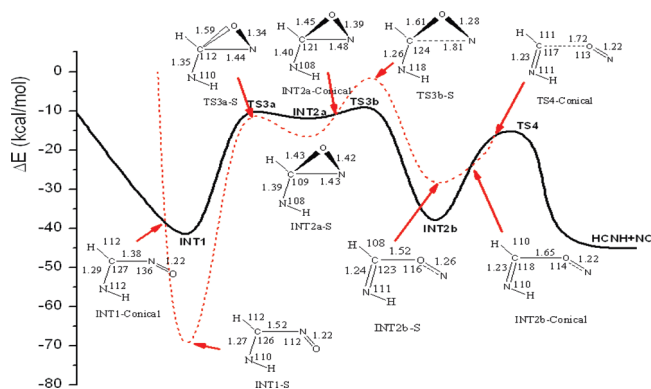


Figure 4. Singlet–triplet crossings on the potential-energy surface of the NH + HCNO → HCNH + NO reaction. Solid line: triplet. Dashed line: singlet. The energies (in kcal/mol) were calculated at the state-averaged MCSCF(14e,12o)/6-311+G(d,p) level of theory with respect to the reactant asymptote. The geometries of the singlet species and the conical interactions are shown. Bond distances are in Ångstroms, and angles are in degrees.

cal interactions between triplet and singlet surfaces relevant to the NH + HCNO → HCNH + NO reaction have been studied. The energetic profiles are shown in Figure 4.

The open-shell singlet INT1-S was first located as shown in Figure 4. There are two differences between INT1 and INT1-S, namely, the CN bond length and the CNO bond angle. INT1-S has a longer CN bond but much smaller CNO angle. In terms of energy, INT1-S is much more stable than the triplet INT1. The conical interaction between INT1-S and INT1 is shown as INT1-conical in Figure 4. It has the same CN bond length with INT1, and its energy is close to INT1 as well. Apparently, the S–T crossing might occur through changing the CNO angles.

Through the potential-energy scan of the CN bond distances it is found that the CN bond fission of INT1-S is barrierless, leading to HCNH + NO directly without any crossing with the triplet path. The O-migration mechanism is complex. As shown in Figure 4, the first part of the reaction path involves singlet TS3a-S and INT2a-S. Later, there is a crossing between INT2a and INT2a-S via INT2a-conical, whose geometry and energy are nearly the same as INT2a. TS3b-S is a much higher barrier, and the breaking C...N bond is much longer than that of TS3b. The intermediate INT2b-S has a similar geometry with INT2b except for the longer CO bond. As the CO bond is stretching, there are two crossing intersections between triplet and singlet, namely, INT2b-conical and TS4-conical. However, we cannot obtain reasonable energetic profiles on the singlet surface when the CO distances are elongated further because the MCSCF calculations cannot converge to the correct electronic states. Probably, the dissociation asymptote of INT2b corresponds to the electronically excited products.

At present, we are unable to calculate the singlet–triplet transition probability. However, it is probably small as is typical of a spin forbidden transition in the absence of heavy elements. The most energetically favorable reaction path, namely, NH + HCNO → TS1 → INT1 → TS2 → HCNH + NO, could be a reasonable mechanistic representation for

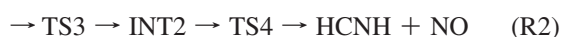
the NH + HCNO reaction and will be used for the subsequent kinetic calculation.

5. H-Shift Reaction Paths from INT1. There are a few H-shift reaction pathways starting from either INT1 or INT2b via five-, four-, or three-membered ring transition states. Although all these channels involve significant barriers, the energies of transition states are still lower than that of the reactant asymptote. It is conceivable that the 1,4-H-shift path has the lowest barrier, forming INT3 via transition state TS5. Subsequently, the C–N bond cleavage produces HCN and HON. It is worth noting that HON is in its excited triplet state rather than the singlet because of the spin conservation.

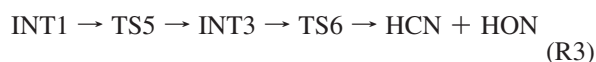
Transition state TS7 is for the four-center 1,3-H-shift path, leading to INT4. The energy of TS7 is more than 10 kcal/mol higher than that of TS5. However, INT4 is more stable than INT3. The CN bond fission leads to HCN and triplet HNO via TS8. It is noted that HCN + HNO is the most exothermic channel for the NH + HCNO reaction, even though HNO is produced in the triplet excited state. The other four-center transition state is TS9, showing a H-shift pathway from C to O. The intermediate INT5 can decompose to HNC and HON via TS10. Another H-shift starting from INT1 occurs via TS11, which is a three-center structure. The energy of TS11 is about 10 kcal/mol higher than that of the NH + HCNO asymptote. Therefore, this H-shift path could be unimportant, although intermediate INT6 is very stable in view of its low energy and the decomposition barrier (TS12) for the HNC + HNO channel is relatively low.

There are two H-shift pathways starting from INT2b. 1,4-N,N shift occurs via TS13, whose energy is about 10 kcal/mol higher than that for the decomposition of INT2b. The intermediate INT7 decomposes to HCN and HNO via TS14. The other H-shift path involves a more significant barrier since TS15 is a four-center transition state. However, the intermediate INT8 still lies below the NH + HCNO asymptote. The CO bond cleavage of INT8 produces HNC and HNO via TS16. The NO bond stretching of INT8 leads to NH and HCNO, which is the most stable species in the [H,C,N,O] system.

Although there are many low-energy reaction pathways in the NH + HCNO reaction, the barriers are well separated. In conclusion, it appears to have a relatively simple reaction mechanism. The major mechanism can be simplified as follows:



Evidently, HCNH and NO should be the major products of the reaction. Another channel starting from INT1, namely,



could be a very minor path under the specific conditions.

6. Temperature- and Pressure-Dependent Rate Constants. For the sake of completeness, the whole mechanism as shown in Figure 2 has been employed for the kinetic analysis using the conventional transition state theory and

the multichannel Rice–Ramsperger–Kassel–Marcus theory (RRKM). The algebraic expressions of the RRKM calculations have been well developed by us and successfully used for kinetic estimates of many radical reactions.^{41,42} Formulas and parameters in the kinetic calculation are deposited in the Supporting Information (see Note S1) in detail. Briefly, the rigid rotor harmonic oscillator (RRHO) approximation was used for all stationary points at the *E*-resolved level. The transition states were treated nonvariationally, and the tunneling effect has been neglected. In the calculation of the pressure dependence, N₂ was used as the bath gas using Troe's weak-collision model.⁴³ It is worth noting that the present rate constants are only rough estimates in view of the uncertainty of the theoretical barrier heights and various approximations used in the transition state theory and master equation analysis.

Since both RS2C/CBS//MCSCF(14e,12o)/6-311+G(d,p) and FCC/CBS//RCCSD(T,full)/6-311+G(d,p) methods give similar energetic and geometrical data for the key stationary points (e.g., the intermediates and transition states in channels R1–R3), the ab initio data calculated at the RS2C/CBS//MCSCF(14e,12o)/6-311+G(d,p) level of theory were used to predict the rate constants as a function of temperature. It is worth noting that HCNO was considered to be a collinear species.

It appears that the NH + HCNO reaction is pressure independent. The association reaction is highly exothermic. The reaction well INT1 is as deep as 61 kcal/mol, and the decomposition of INT1 to form HCNH and NO is a nearly barrierless process. This feature of the potential-energy surface implies that the energy-rich intermediate INT1 has little chance to be quenched by the bath gases except for at extremely high pressures. Our transition state theory calculation does verify this fact (e.g., $P > 10^6$ Torr of bath gas is required to produce significant INT1).

The temperature-dependent rate constants for the NH + HCNO reaction in the temperature range 250–3000 K are shown in Figure 5. The HCNH + NO channel is always the dominant path, and the other product channels contribute less than 1%. The overall rate constants show positive temperature dependence in accordance with the small barrier at the entrance. Moreover, a typical non-Arrhenius curvature is observed. The rate constants were least-squares fitted using an empirical three-parameter expression as follows

$$k(T = 250 - 3000 \text{ K}) = 1.81 \times 10^{-13} (T/300)^{1.97} e^{-1.39 \text{ kcal/mol}/RT} \quad (7)$$

in units of cm³ molecule⁻¹ s⁻¹. As shown in Figure 5, the fitting to the low- and high-temperature ranges separately shows better result, namely,

$$k(T < 1000 \text{ K}) = 3.56 \times 10^{-13} (T/300)^{1.67} e^{-1.98 \text{ kcal/mol}/RT} \quad (8)$$

$$k(T > 1000 \text{ K}) = 1.72 \times 10^{-13} (T/300)^{1.98} e^{-1.32 \text{ kcal/mol}/RT} \quad (9)$$

At lower temperatures, the NH + HCNO reaction occurs slowly. Under the high-temperature combustion conditions, the reaction becomes fast. For example, the rate constant is

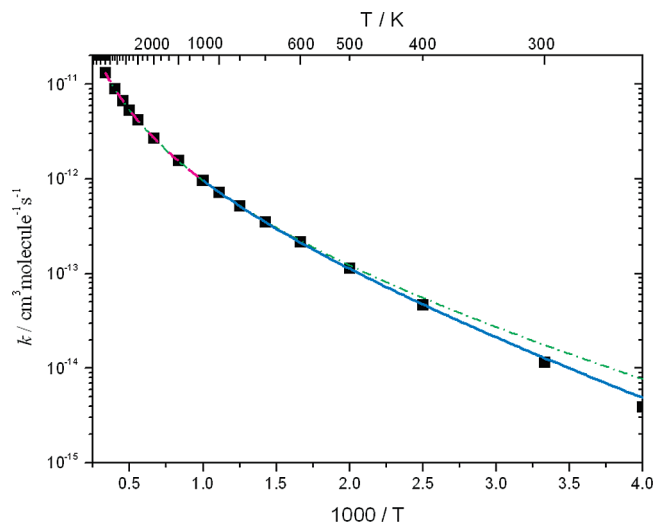


Figure 5. Overall rate constants for the NH + HCNO reaction as a function of temperatures calculated using the RS2C/CBS//MCSCF(14e,12o)/6-311+G(d,p) ab initio data. Solid line: best fit in the range 250–1000 K. Dashed line: best fit in the range 1000–3000 K. Dash-dotted line: best fit to the entire data.

$5.33 \times 10^{-12} \text{ cm}^3 \text{ molecule}^{-1} \text{ s}^{-1}$ at 2000 K and increases to be twice that at 3000 K. More significantly, the major reaction products include both HCNH and NO. It implies that the NH + HCNO reaction might be a potentially important source of NO regeneration under the combustion conditions. On the other hand, decomposition of the HCNH radical into either $\text{H} + \text{HCN}$ or $\text{H}_2 + \text{CN}$ is also energetically feasible because of the significant energy released in the $\text{HCNH} + \text{NO}$ channel.^{44,45} Experimental study of the NH + HCNO reaction might be interesting in view of the present theoretical findings.

IV. Conclusions

In this work, two types of high-level ab initio methods, namely, multireference Rayleigh–Schrodinger perturbation theory based on the multiconfigurational self-consistent-field theory and the coupled cluster theory along with the complete basis set extrapolations, were used to explore the mechanism of the NH + HCNO reaction. On the triplet potential-energy surface, a small barrier of about 3 kcal/mol exists for the $\text{C}\cdots\text{N}$ association path. The association step is highly exothermic by more than 60 kcal/mol, and the intermediate $\text{HC}(\text{NH})\text{NO}$ can decompose rapidly to form HCNH and NO radicals, which are the dominant products. Another two routes for the $\text{HCNH} + \text{NO}$ production include the symmetry-breaking O-rearrangement process and the singlet–triplet surface crossing path. Other products such as HCN and HNO are shown to be unimportant.

The rate constants for the NH + HCNO reaction were predicted on the basis of the ab initio data. The rate constants are positively temperature dependent and show non-Arrhenius behavior. It is shown that the NH + HCNO reaction is fast under high-temperature combustion conditions. Moreover, the rapid production of NO and HCNH (may decompose to $\text{H} + \text{HCN}$ or $\text{CN} + \text{H}_2$) implies that the NH +

HCNO reaction might play an important role in the NO-reburning process.

Acknowledgment. This work was supported by NSFC (no. 20673079, 20603025).

Supporting Information Available: Figures of the active spaces in the MCSCF calculations for some key species involved in the NH + HCNO reaction and G4//MP2(full)/6-311+G(d,p) calculated potential-energy surface for the NH + HCNO reaction. Tables of the MCSCF(14e, 12o)/6-311+G(d,p) and RCCSD(T,full)/6-311+G(d,p) calculated harmonic vibrational frequencies and optimized geometries for TS1 and INT1 with different basis sets. Note S1 lists the details for the kinetic calculation. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Beck, D. W.; Feldl, K. *Angew. Chem., Int. Ed.* **1966**, *5*, 722.
- (2) Schuurman, M. S.; Muir, S. R.; Allen, W. D.; Schaefer, H. F., III. *J. Chem. Phys.* **2004**, *120*, 11587.
- (3) Nguyen, M. T.; Pierloot, K.; Vanquickenborne, L. G. *Chem. Phys. Lett.* **1991**, *181*, 83.
- (4) Koput, J.; Winnewisser, B. P.; Winnewisser, M. *Chem. Phys. Lett.* **1996**, *255*, 357.
- (5) Zhang, W. C.; Du, B. N.; Feng, C. J. *Chem. Phys. Lett.* **2007**, *442*, 1.
- (6) Li, B. T.; Zhang, J.; Wu, H. S.; Sun, G. D. *J. Phys. Chem. A* **2007**, *111*, 7211.
- (7) Feng, W. H.; Hershberger, J. F. *J. Phys. Chem. A* **2007**, *111*, 3831.
- (8) Feng, W. H.; Hershberger, J. F. *J. Phys. Chem. A* **2006**, *110*, 12184.
- (9) Grimme, S.; Muck-Lichtenfeld, C.; Wurthwein, E. U.; Ehlers, A. W.; Goumans, T. M.; Lammertsma, K. *J. Phys. Chem. A* **2006**, *110*, 2583.
- (10) Feng, W. H.; Hershberger, J. F. *Chem. Phys. Lett.* **2008**, *457*, 307.
- (11) Miller, J. A.; Klippenstein, S. J.; Glarborg, P. *Combust. Flame* **2003**, *135*, 357.
- (12) Marcelino, N.; Cernicharo, J.; Tercero, B.; Roueff, E. *Astrophys. J.* **2009**, *690*, L27.
- (13) Quandt, R. W.; Hershberger, J. F. *J. Phys. Chem.* **1995**, *99*, 16939.
- (14) Romming, H. J.; Wagner, H. G. *Symp. Int. Combust. Proc.* **1996**, *26*, 559.
- (15) Mullen, C.; Smith, M. A. *J. Phys. Chem. A* **2005**, *109*, 1391.
- (16) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- (17) Moller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618.
- (18) Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. *J. Chem. Phys.* **1980**, *72*, 650.
- (19) 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. *Gaussian03*, Revision D.01; Gaussian, Inc.: Wallingford, CT, 2004.
- (20) Werner, H.-J.; Knowles, P. J. *J. Chem. Phys.* **1985**, *82*, 5053.
- (21) Knowles, P. J.; Werner, H.-J. *Chem. Phys. Lett.* **1985**, *115*, 259.
- (22) Knowles, P. J.; Hampel, C.; Werner, H.-J. *J. Chem. Phys.* **1993**, *99*, 5219 Erratum: Knowles, P. J.; Hampel, C.; Werner, H.-J. *J. Chem. Phys.* **2000**, *112*, 3106.
- (23) Lee, T. J.; Taylor, P. R. *Int. J. Quantum Chem. Symp.* **1989**, *23*, 199.
- (24) Busch, T.; Degli-Esposti, A.; Werner, H.-J. *J. Chem. Phys.* **1991**, *94*, 6708.
- (25) Eckert, F.; Pulay, P.; Werner, H.-J. *J. Comput. Chem.* **1997**, *18*, 1473.
- (26) Rauhut, G.; El-Azhary, A.; Eckert, F.; Schumann, U.; Werner, H.-J. *Spectrochim. Acta* **1999**, *55*, 651.
- (27) Hrenar, T.; Rauhut, G.; Werner, H.-J. *J. Phys. Chem. A* **2006**, *110*, 2060.
- (28) Eckert, F.; Werner, H.-J. *Theor. Chem. Acc.* **1998**, *100*, 21.
- (29) Werner, H.-J. *Mol. Phys.* **1996**, *89*, 645.
- (30) Celani, P.; Werner, H.-J. *J. Chem. Phys.* **2000**, *112*, 5546.
- (31) Roos, B. O.; Andersson, K. *Chem. Phys. Lett.* **1995**, *245*, 215.
- (32) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007.
- (33) Helgaker, T.; Klopper, W.; Koch, H.; Noga, J. *J. Chem. Phys.* **1997**, *106*, 9639.
- (34) Feller, D. *J. Chem. Phys.* **1993**, *98*, 7059.
- (35) He, H.; He, Z.; Cremer, D. *Chem. Phys. Lett.* **2000**, *317*, 535.
- (36) Truhlar, D. G. *Chem. Phys. Lett.* **1998**, *294*, 45.
- (37) Werner, H.-J.; Knowles, P. J.; Lindh, R.; Manby, F. R.; Schütz, M.; Celani, P.; Korona, T.; Mitrushenkov, A.; Rauhut, G.; Adler, T. B.; Amos, R. D.; Bernhardsson, A.; Berning, A.; Cooper, D. L.; Deegan, M. J. O.; Dobbyn, A. J.; Eckert, F.; Goll, E.; Hampel, C.; Hetzer, G.; Hrenar, T.; Knizia, G.; Koppl, C.; Liu, Y.; Lloyd, A. W.; Mata, R. A.; May, A. J.; McNicholas, S. J.; Meyer, W.; Mura, M. E.; Nicklass, A.; Palmieri, P.; Pflüger, K.; Pitzer, R.; Reiher, M.; Schumann, U.; Stoll, H.; Stone, A. J.; Trarroni, R.; Thorsteinsson, T.; Wang, M.; Wolf, A. *MOLPRO2006.1, A package of ab initio programs*; University College Cardiff Consultants Ltd: Cardiff, U.K., 2006; <http://www.molpro.net>.
- (38) McDouall, J. J.; Peasley, K.; Robb, M. A. *Chem. Phys. Lett.* **1988**, *148*, 183.
- (39) Werner, H.-J.; Knowles, P. J. *J. Chem. Phys.* **1988**, *89*, 5803.
- (40) Knowles, P. J.; Werner, H.-J. *Chem. Phys. Lett.* **1988**, *145*, 514.
- (41) Hou, H.; Wang, B. *J. Chem. Phys.* **2007**, *127*, 054306.
- (42) Hou, H.; Li, A.; Hu, H.; Li, Y.; Li, H.; Wang, B. *J. Chem. Phys.* **2005**, *112*, 224304.
- (43) Troe, J. *J. Chem. Phys.* **1977**, *66*, 4745.
- (44) Metropoulos, A.; Thompson, D. L. *J. Mol. Struct: THEOCHEM* **2007**, *822*, 125.
- (45) Horst, M. A.; Schatz, G. C.; Harding, L. B. *J. Chem. Phys.* **1996**, *105*, 558.

CT9002288

Topology of the Effectively Paired and Unpaired Electron Densities for Complex Bonding Patterns: The Three-Center Two-Electron Bonding Case

Rosana M. Lobayan,[†] Roberto C. Boichichio,^{*,‡} Alicia Torre,[§] and Luis Lain[§]

Facultad de Ingeniería, Universidad de la Cuenca del Plata, Lavalle 50, 3400 Corrientes, Argentina, Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Ciudad Universitaria, 1428 Buenos Aires, Argentina, and Departamento de Química Física, Facultad de Ciencias, Universidad del País Vasco, Apdo, 644 E-48080 Bilbao, Spain

Received June 2, 2009

Abstract: Our previously reported local formalism of the electron density decomposition into effectively paired and unpaired densities is applied to electron deficient molecular systems possessing complex bonding patterns. It is shown that the unpaired density is not only near the nuclear positions, like in classical bonds, but also spills out over the bonding regions, to compensate the electron deficiency. Topological information obtained from the effectively unpaired density, which may not be directly observed from the total density, allows us to establish a procedure to detect complex interactions. This study is complemented with results arising from nonlocal formalism of topological population analyses. The conclusions from both formalisms are in complete agreement and permit to interpret the well-known structural information from Lipscomb *styx* numbers going beyond it in cases where the electronic description becomes ambiguous, pointing out the subtle information contained in the unpaired density. Numerical results for three-center two-electron bondings in the boranes B₂H₆, B₄H₁₀, B₅H₉, and B₅H₁₁ are reported.

1. Introduction

The physical information contained in N -electron molecular state functions is usually summarized by means of chemical concepts, such as atomic and bondings populations, bond orders^{1–4} and ionicities (multiplicities),^{4–6} atomic and free valencies,^{4,7–12} etc. In the quantitative description of these concepts the electron density^{13–16} constitutes a tool of paramount interest. Two complementary methodologies have been developed to deal with this density. On the one hand the *nonlocal or integrated formulations*, which are known as electronic population analysis.^{4,13–16} Within these methods, the Mulliken-type partitionings¹⁷ associate an atom with the atomic basis functions centered on the corresponding

nucleus in the molecule; the atomic and bonding regions become automatically defined, regarding the localization of these atomic functions.^{18,19} Other methods of population analyses assign real space regions to the atoms, as the “fuzzy” atoms descriptions,^{20–23} projection operators techniques²⁴ or the topological atoms in molecules (AIM) theory.^{7,25–30} On the other hand, the electron density $\rho(\mathbf{r})$ and its associated Laplacian field $\nabla^2\rho(\mathbf{r})$ can be studied through their topological structures. The localization and classification of their critical points (cp), i.e., maxima, minima, or saddle ones, characterize these fields, providing an alternative and complementary approach to extract and interpret chemical information.^{25,26} This approach will be called *local formulation*.

The cp's are classified regarding the value and sign of the three eigenvalues of the Hessian matrix of the density $\rho(\mathbf{r})$. This information is featured as (r, s) , where r is the *rank* (the number of nonzero eigenvalues of the Hessian matrix) and s is the *signature* (the sum of the signs of its eigenval-

* Corresponding author e-mail: rboc@df.uba.ar.

[†] Universidad de la Cuenca del Plata.

[‡] Universidad de Buenos Aires.

[§] Universidad del País Vasco.

ues). Any point with all negative eigenvalues is denoted as (3,-3); such a point is called a nuclear critical point (ncp) and indicates a local maximum placed at the nuclear positions. A cp with two negative eigenvalues and a positive one is denoted by (3,-1) and corresponds to a bond critical point (bcp); in this type of cp's the concentration of the electron density shows a bonding interaction between two atoms.^{25,26} The first two eigenvalues of the Hessian matrix correspond to the perpendicular curvatures, and the third one provides a curvature along the internuclear axis. In this scenario, a covalent bond is featured by the electron cloud possessing two large negative curvatures perpendicular to the bond line and a small positive curvature along the bond at the position of the bcp.^{25,26} Other important cp's which are not present in systems with classical bonding patterns are the *ring* (rcp) and *cage* (ccp) critical points, which appear in molecular systems of complex electronic structure; these points are characterized by rank and signature (3,+1) and (3,+3), respectively.^{25,26} The value of $\nabla^2\rho(\mathbf{r})$ is the sum of the curvatures along the orthogonal coordinate axes. Its sign indicates electron density locally depleted (positive) or locally concentrated (negative) and constitutes valuable information to describe the behavior of the density around a local point^{25,26,31} and other features of the electron distribution, as the unpaired electron contributions to the density.^{32,33}

In previous works^{32,33} we have reported an exact partitioning of the electron density so that $\rho(\mathbf{r}) = \rho^{(p)}(\mathbf{r}) + \rho^{(u)}(\mathbf{r})$ in which the contributions $\rho^{(p)}(\mathbf{r})$ and $\rho^{(u)}(\mathbf{r})$, possessing different physical character, correspond to the effectively paired and unpaired electron density, respectively.⁸⁻¹¹ We have described the behavior of both contributions to the electron density and the shifts of their (3,-3) and (3,-1)cp's in comparison with those of the total density. Our results show that the positions of the (3,-3)cp's of $\rho^{(p)}(\mathbf{r})$ are located very close to the total $\rho(\mathbf{r})$ ncp, while the $\rho^{(u)}(\mathbf{r})$ (3,-3)cp's are also located close to the nuclear positions but out of the bonding region.³² The utilization of the Laplacian functions, $\nabla^2\rho^{(p)}(\mathbf{r})$ and $\nabla^2\rho^{(u)}(\mathbf{r})$, shows that nuclear regions possess most of the effectively unpaired electrons. Moreover, both paired and unpaired density fields possess successive regions of concentration and depletion of density, yielding a shell structure.³³ These results have been obtained in systems possessing conventional patterns of bonding. Hence, we have now the challenge to extend this treatment to systems with more complex bonding patterns, such as three-center two-electron (3c-2e) bonds. For this goal, we have chosen several electron deficient boron hydrides, due to the crucial importance of the presence of three-center bondings in the structural description of these compounds.³⁴⁻³⁷

The organization of this article is as follows. The second section briefly reports the theoretical framework of the partitioning of the electron density and the relationships between the density gradients and between the Laplacian fields as well as the tools used to carry out complementary studies of topological population analysis. The third section describes the computational details of the calculations performed over a set of selected boranes and a comparative discussion of the results from both formulations. Finally, the last section is devoted to the concluding remarks.

2. Theoretical Background

The electron density, $\rho(\mathbf{r})$, in an N -electron molecular system may be decomposed into two contributions³²

$$\rho(\mathbf{r}) = \rho^{(p)}(\mathbf{r}) + \rho^{(u)}(\mathbf{r}) \quad (1)$$

where $\rho^{(p)}(\mathbf{r})$ and $\rho^{(u)}(\mathbf{r})$ stand for the effectively paired and unpaired densities, defined by

$$\rho^{(p)}(\mathbf{r}) = \frac{1}{2} \int d\mathbf{r}' {}^1D(\mathbf{r}|\mathbf{r}') {}^1D(\mathbf{r}'|\mathbf{r}) \quad (2)$$

and

$$\rho^{(u)}(\mathbf{r}) = \frac{1}{2} u(\mathbf{r}|\mathbf{r}) \quad (3)$$

respectively. ${}^1D(\mathbf{r}|\mathbf{r}')$ means the spin-free first-order reduced density matrix (1-RDM) in the coordinate representation,^{14,38} its trace (coordinate integration over the whole real space) is the number of electrons in the system, i.e., $tr({}^1D) = \int d\mathbf{r} {}^1D(\mathbf{r}|\mathbf{r}) = \int d\mathbf{r} \rho(\mathbf{r}) = N$. $u(\mathbf{r}|\mathbf{r})$ is the diagonal element of the effectively unpaired density matrix defined by $u(\mathbf{r}|\mathbf{r}') = 2 {}^1D(\mathbf{r}|\mathbf{r}') - {}^1D^2(\mathbf{r}|\mathbf{r}')$ where ${}^1D^2(\mathbf{r}|\mathbf{r}') = \int d\mathbf{r}'' {}^1D(\mathbf{r}|\mathbf{r}'') {}^1D(\mathbf{r}''|\mathbf{r}')$.⁸⁻¹² Note that the densities are defined as the diagonal part of the corresponding density matrices, e.g. $\rho(\mathbf{r}) = {}^1D(\mathbf{r}|\mathbf{r})$. The physical meaning of the traces of the effectively paired and unpaired densities is the number of paired (opposite spins) and unpaired (far away) electrons. The unpaired density has two sources, one of them comes from the spin density (only present in nonsinglet states) and the other corresponds to the many-body effects or correlation effects that are supported by the Coulomb interaction between the particles.^{7,8,28} It may be noted that for state functions having all orbitals doubly occupied, as in the closed-shell Hartree-Fock or the density functional theory (DFT) cases, $\rho^{(u)}(\mathbf{r})$ is intrinsically zero,³² and such effects cannot be detected.

The techniques used to study the electron density topology^{25,26} may also be applied to its two above-mentioned components, to describe bonding features in a more precise manner. The fundamental information is obtained from the localization of the cp's of each contribution of the electron density. The cp's for the total density are found throughout the gradient of the field by

$$\nabla\rho(\mathbf{r}) \Big|_{\mathbf{r}^c} = 0 \quad (4)$$

or equivalently, according to eq 1

$$\nabla\rho^{(p)}(\mathbf{r}) \Big|_{\mathbf{r}^c} + \nabla\rho^{(u)}(\mathbf{r}) \Big|_{\mathbf{r}^c} = 0 \quad (5)$$

where $\mathbf{r}^c = \{\mathbf{r}_i^c; i = 1, \dots, M\}$ indicates the set of critical points of the total density $\rho(\mathbf{r})$. Then the following relation holds

$$\nabla\rho^{(p)}(\mathbf{r}) \Big|_{\mathbf{r}^c} = -\nabla\rho^{(u)}(\mathbf{r}) \Big|_{\mathbf{r}^c} \quad (6)$$

whose physical meaning is that each density field component increases/decreases its value in opposite direction, i.e., when

$\rho^{(p)}(\mathbf{r})$ increases its value, the other part of the density, the unpaired contribution, properly does it in the opposite direction. Hence, no simultaneous increment/decrease of each one may occur at the cp's.³² The Laplacian field of eq 1 yields

$$\nabla^2 \rho(\mathbf{r}) \Big|_{\mathbf{r}^c} = \nabla^2 \rho^{(p)}(\mathbf{r}) \Big|_{\mathbf{r}^c} + \nabla^2 \rho^{(u)}(\mathbf{r}) \Big|_{\mathbf{r}^c} \neq 0 \quad (7)$$

According to eq 7 both $\nabla^2 \rho^{(p)}(\mathbf{r}) \Big|_{\mathbf{r}^c}$ and $\nabla^2 \rho^{(u)}(\mathbf{r}) \Big|_{\mathbf{r}^c}$ contributions do not necessarily follow opposite trends, as the above-mentioned ones for the density gradient. Thus, both density Laplacians may concentrate or deplete at the neighborhood of a cp simultaneously. It is worthwhile to note that the results reported in ref 33 have shown that both $\rho^{(p)}(\mathbf{r})$ and $\rho^{(u)}(\mathbf{r})$ densities possess a shell structure as the $\rho(\mathbf{r})$ one.^{33,39}

The nonlocal or integrated formalism is complementary to the above-mentioned local one. This nonlocal formalism deals with classical chemical concepts like atomic charges, bond orders, and valences, etc.^{7,8,28–30} We will refer to the nonlocal AIM topological population analysis. We summarize the relations defining the relevant magnitudes to our goal, which have been previously stated within that approach.^{7,29} The two-center bond index has been defined as

$$I_{\Omega_A \Omega_B} = \sum_{i,j,k,l} {}^1D_j^i {}^1D_l^k S_{il}(\Omega_A) S_{kj}(\Omega_B) \quad (8)$$

where Ω_A and Ω_B stand for Bader's atomic domains in the physical space,²⁵ ${}^1D_j^i$ are the first-order reduced density matrix elements, and $S_{ij}(\Omega_A)$ are the elements of the overlap matrix over the region Ω_A , in the orthogonal molecular basis set $\{i,j,k,l,\dots\}$.⁷ In the same context, the three-center bond index is

$$I_{\Omega_A \Omega_B \Omega_C} = \sum_{i,j,k,l,m,n} {}^1D_j^i {}^1D_l^k {}^1D_n^m S_{in}(\Omega_A) S_{kj}(\Omega_B) S_{ml}(\Omega_C) \quad (9)$$

giving rise to the three-center topological bond order as²⁹

$$\Delta_{\Omega_A \Omega_B \Omega_C}^{(3)} = \frac{1}{4} \sum_{P(\Omega_A \Omega_B \Omega_C)} I_{\Omega_A \Omega_B \Omega_C} \quad (10)$$

where $P(\Omega_A \Omega_B \Omega_C)$ indicates the permutations of the three domain contributions.

The theoretical structural index or *styx* number defined by Lipscomb is the indicator commonly used to characterize the reported compounds.³⁷ According to Lipscomb rules the structure of electron deficient boranes can be summarized by this index where *s* is the number of the 3-center BHB bonds, *t* is the number of 3-center BBB bonds, *y* is the number of 2-center BB bonds, and *x* gives the number of BH₂ groups in the system.

As has been pointed out in the Introduction, the main purpose of this paper is to study the capability of these tools to describe nonconventional bondings as the (3c-2e) ones. The results are reported in the next section.

3. Results and Discussion

3.1. Computational Details and Results. The state functions used in this work to describe the molecular systems were calculated at the level of configuration interaction with single and double excitations (CISD), using the Gaussian 03 package⁴⁰ with the basis sets 6-31G**. The geometries for all systems were optimized within this approximation. The densities, their critical points, and their Laplacian fields $\nabla^2 \rho^{(p)}(\mathbf{r})$ and $\nabla^2 \rho^{(u)}(\mathbf{r})$ were determined by appropriately modified AIMPAC modules.⁴¹ The reported systems have been chosen to study the (3c-2e) complex pattern of bonding present in electron deficient boranes^{34–36} (some preliminary results for the B₂H₆ system have been reported in ref 32). The numerical results of populations analysis were obtained with our own codes mentioned in ref 29. For practical reasons, we will use the function $L(\mathbf{r}) = -\nabla^2 \rho(\mathbf{r})$ in the discussion of results as an indicator of concentration (positive value) or depletion (negative value) of the number of electrons at the point \mathbf{r} ;^{32,33} the terms accumulation and reduction have been proposed for the description of maxima and minima in $\rho(\mathbf{r})$.^{39,42} Because of the complex structure of the $\rho^{(u)}(\mathbf{r})$ topology, we will only deal with critical points associated with its valence shells (vs) in the corresponding systems, and no reference will be made to those of the inner shells of this density; in fact, only the former ones are involved in bonding phenomena. The terminology vs(3,-1)cp, vs(3,+1)cp, and vs(3,+3)cp will refer to (3,-1), (3,+1), and (3,+3) critical points of $\rho^{(u)}(\mathbf{r})$ valence shell, in analogy with the bcp, rcp, and ccp's of the total density. Nevertheless, it is important to note that such points are not *sensu strictu* bcp, rcp, or ccp's because only the cp's of the total density are able to define a bond in the AIM topological formalism.^{25,26} Figure 1 shows the molecular structure of each of the systems studied in this work.

Table 1 shows a detailed description of the electronic distributions from the topological (local and nonlocal) point of view for these borane systems. We will use diborane molecule, B₂H₆, as a test example to show the similar and complementary description of the electron distribution provided by the local and nonlocal formalisms. Table 2 shows the numerical values for each density and its $L(\mathbf{r})$ field at the bcp's of the total density and at $\rho^{(u)}(\mathbf{r})$ vs(3,-1) cp's; Tables S1 and S2 of the Supporting Information (SI) show the numerical values for the ncp's and rcp's, respectively.

B₂H₆. As is indicated in Table 1, the structural number *styx* for this hydride is 2002. It means that this system possesses two (3c-2e) bridge bonds BHB, no BBB bonds, no BB bonds, and two groups BH₂. The topological structure of the total density ρ exhibits one ncp located on each nucleus, one bcp for each of the four terminal bond BH, and a bcp defining a bond for each one of the four bridge sequences BH. The absence of bcp's between the boron atoms indicates that this topology does not detect a bond BB in this system. Nevertheless, a rcp defined by the sequence BHBH appears, which is located at the central point of the line joining the boron atoms. These features are coincident with the structural information provided by the

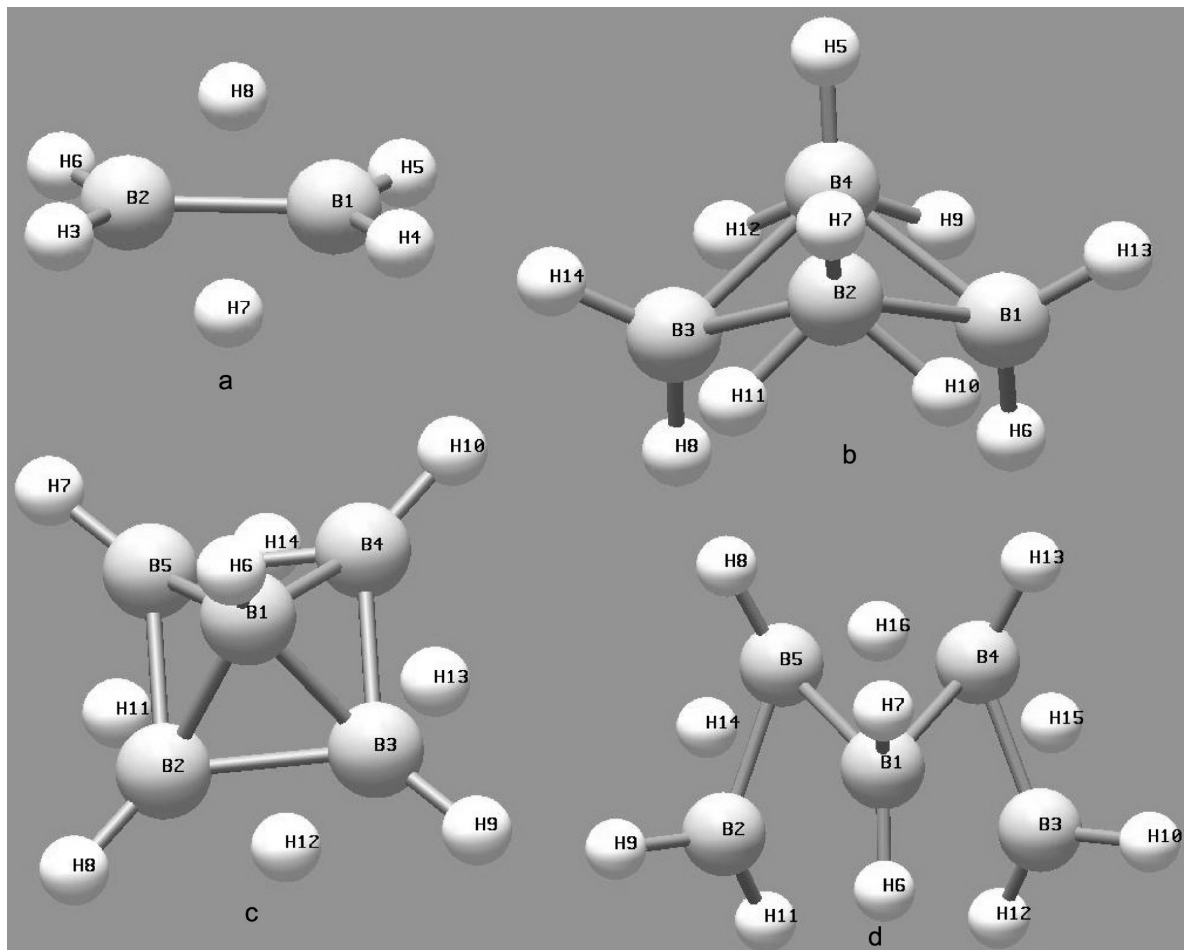


Figure 1. Geometrical conformation of borane systems: (a) B_2H_6 , (b) B_4H_{10} , (c) B_5H_9 , and (d) B_5H_{11} . Atomic symbols and labels are shown.

number *styx* except for the two bridge bonds BHB, pointed out by its 2002 value, i.e., the bridge BH bcp's and the BHBH rcp, which indicates an intersection point among the BH bridge atomic basins, reveal interactions among all these basins but do not provide enough information to distinguish between two three-center bondings or one four-center bonding involving the atoms forming the ring. The structure of the effectively paired density $\rho^{(p)}$ is similar to that of the total one ρ , as may be observed in Table 1, and therefore it cannot introduce any new information about the electronic structure.

Let us now inspect the topology of the effectively unpaired density $\rho^{(u)}$. The structure of this density is similar to those of the above commented ρ and $\rho^{(p)}$ ones, but it presents some additional subtle and crucial features. As may be observed in Table 1, a $\rho^{(u)}$ vs(3,-1)cp appears for the sequence BB which is not present for ρ and $\rho^{(p)}$, having the same location than the unique (3,+1)cp of these last densities (see Figure 2). Moreover two vs(3,+1)cp's are detected, one for each of the two bridge bond BHB, which are symmetrically placed on the line joining the bridge H atoms at both sides of the BB vs(3,-1)cp. These features may clearly induce to interpret such new elements as indicators of two (3c-2e) bonds. In view of these considerations let us state *in advance* a rule based on the onset of a vs(3,-1)cp between the boron atoms closing

the bridge sequence BHB and a vs(3,+1)cp associated with that sequence: *in borane systems a (3c-2e) bond between atoms ABC exists if there is a vs(3,-1)cp of $\rho^{(u)}$ between each pair of atoms involved in the three-center ABC sequence and a vs(3,+1)cp defined only by the atoms involved in the three center bond*, hereafter called the *local rule*. This result seems to be the local version of the criterion of the integrated formalism of population analysis for detecting three-center bonding, hereafter: *the integrated or nonlocal rule*. That criterion settles down the existence of a (3c-2e) bond between atoms ABC when fractional bond orders $I_{\Omega_A\Omega_B}$ appear between all possible pairs of atoms AB, BC, and AC.^{4,43} As is shown in Table 1, the values of $I_{BB} = 0.056$, bridge $I_{BH} = 0.280$, and $\Delta_{BHB} = 0.061$ indicate the rule is fulfilled. Hence, the information contained in $\rho^{(u)}$ permits to fit the structural and local description of the electron distribution, i.e., both formulations have comparable criteria to establish the existence of a (3c-2e) bond. Consequently, the results for this emblematic system show that the local description of the electron distribution is completely coincident with the integrated one as well as with the structural *styx* number, taking into account the information contained in the effectively unpaired electron density. Let us note that the value for the terminal $I_{BH} = 0.495$ reveals a normal BH

Table 1. Structural *styx* Numbers and Local and Integrated (Nonlocal) Topological Features of $\rho(\mathbf{r})$, $\rho^{(l)}(\mathbf{r})$, and $\rho^{(u)}(\mathbf{r})$ Densities for Boranes Systems at the CISD/6-31G** Level of Approximation^c

cp type	ρ sequences ^a	$\rho^{(l)}$ sequences ^b	$\rho^{(u)}$ sequences ^b	bond	$I_{\Omega_A\Omega_B}$	$\Delta_{\Omega_A\Omega_B\Omega_C}$
<i>B₂H₆ (styx: 2002)</i>						
ncp	one on each B and H atoms	<i>ibid</i> ρ	<i>ibid</i> ρ	BB	0.056	
bcp	one per each BH_{term} bond	<i>ibid</i> ρ	<i>ibid</i> ρ ;	$H_{term}\dots H_{term}$	0.121	
	one per each BH_{bridge} bond		BB	BH_{term}	0.495	
rcp	$BH_{bridge}BH_{bridge}$	<i>ibid</i> ρ	$BH_{bridge}B$ (2)	BH_{bridge}	0.280	
ccp	no cage critical points present	<i>ibid</i> ρ	<i>ibid</i> ρ	$B..H_{bridge}..B$		0.061
				BH_2		0.051
<i>B₂H₈²⁻</i>						
ncp	one on each B and H atoms	<i>ibid</i> ρ	<i>ibid</i> ρ	BB	0.167	
bcp	one per each BH_{term} bond	<i>ibid</i> ρ	<i>ibid</i> ρ	BH_{term}	0.670	
	one per each BH_{bridge} bond			BH_{bridge}	0.442	
rcp	$BH_{bridge}BH_{bridge}$	<i>ibid</i> ρ	<i>ibid</i> ρ	$BH_{bridge}B$		0.024
ccp	no cage critical points present	<i>ibid</i> ρ	<i>ibid</i> ρ	BH_2		0.067
<i>B₄H₁₀ (styx: 4012)</i>						
ncp	one on each B, H atoms	<i>ibid</i> ρ	<i>ibid</i> ρ	B_2B_4	0.476	
bcp	one per each BH_{term} bond	<i>ibid</i> ρ	<i>ibid</i> ρ ;	B_1H_{10}	0.245	
	one per each BH_{bridge} bond		B_1B_2, B_1B_4	B_2H_{10}	0.431	
	B_2B_4			B_1H_{13}	0.492	
rcp	$B_1H_{10}B_2B_4H_9$	<i>ibid</i> ρ	$B_1H_9B_4; B_1B_2B_4$	B_1H_6	0.475	
ccp	no cage critical points present	<i>ibid</i> ρ	<i>ibid</i> ρ	B_2H_7	0.577	
				B_1B_2	0.095	
				$B_1B_2B_4$		0.056
				$B_1H_{10}B_2$		0.086
				$B_1H_9B_4$		0.087
<i>B₅H₉ (styx: 4120)</i>						
ncp	one on each B, H atoms	<i>ibid</i> ρ	<i>ibid</i> ρ	B_1B_2	0.385	
bcp	one per each BH_{term} bond;	<i>ibid</i> ρ	<i>ibid</i> ρ	B_1H_6	0.627	
	one per each BH_{bridge} bond;		one per each $B_{eq}B_{eq}$ bond	B_2H_8	0.601	
	one per each $B_{eq}B_{ax}$ bond			B_2B_5	0.304	
rcp	$B_1H_{10}B_2B_4$	<i>ibid</i> ρ	$B_{eq}B_{ax}B_{eq}; B_2H_{12}B_3$	B_2H_{12}	0.364	
ccp	no cage critical points present	<i>ibid</i> ρ	<i>ibid</i> ρ	B_2H_4	0.081	
				$B_1B_2B_3$		0.138
				$B_1B_2B_4$		0.049
				$B_2H_{12}B_3$		0.118
<i>B₅H₁₁ (styx: 3203)</i>						
ncp	one on each B, H atoms	<i>ibid</i> ρ	<i>ibid</i> ρ	B_1B_2	0.214	
bcp	BH_{bridge} bond	<i>ibid</i> ρ	<i>ibid</i> ρ	B_1B_4	0.459	
	$B_1B_5; B_3H_8; B_2H_9$		$B_1B_2; B_4B_5;$	B_1H_6	0.503	
	$B_2B_5; B_1H_7; B_5H_{14}$		B_2H_{14}	B_1H_7	0.622	
	$B_2H_{11}; B_1H_6$			B_4H_{13}	0.574	
rcp	$B_5B_1B_4H_{bridge}$	<i>ibid</i> ρ	$B_5H_{bridge}B_4; B_5B_1B_4$	B_1H_{16}	0.067	
			$B_5B_1B_2; B_5B_2H_{14}$	B_2B_3	0.012	
				B_2B_5	0.206	
				B_2H_7	0.022	
				B_2H_6	0.166	
				B_3H_{10}	0.524	
				B_3H_{12}	0.496	
				B_3H_{15}	0.264	
				B_3B_4	0.207	
				B_4B_5	0.172	
				B_4H_{15}	0.435	
				B_4H_{16}	0.348	
				B_5H_8	0.571	
				$B_1B_2B_5$		0.137
				$B_1B_2H_6$		0.092
				$B_1B_4B_5$		0.113
				$B_2B_5H_{14}$		0.109
				$B_4B_5H_{16}$		0.105
ccp	no cage critical points present					

^a Indicate the nucleus at which the **ncp** is located; for **bcp**s, the atoms defining the bond; for **rcp**s, the atoms giving rise to the ring; for **ccp**s, the atoms defining the cage. ^b Only those of ρ are *sensu stricto* **cps**. ^c All quantities are in atomic units.

bonding in the terminal groups, and thus their H atoms may not be involved within (3c-2e) bond.

Table S1 collects the results about the local charge distribution over the nuclear attractors (nuclei) showing electronic charge concentration near the nucleus, as expected.

Table 2 reports the numerical results for the densities at the bcp's of the total density (first line) and at the $vs(3,-1)cp$'s of the effectively unpaired density (second line). The three density fields show electron concentration at the total density bcp's for the BH bonds in each BH₂ group, and the same

Table 2. Density and $L(\mathbf{r})$ for the Total Density ρ and Its Contributions $\rho^{(p)}$, $\rho^{(u)}$ at Bond Critical Points of Total Density at the CISD/6-31G** Level of Calculation^{a,e}

system	bond	$\rho(\mathbf{r}) _{bcp}$	$\rho^{(p)}(\mathbf{r}) _{bcp}$	$\rho^{(u)}(\mathbf{r}) _{bcp}$	$-\nabla^2\rho(\mathbf{r}) _{bcp}$	$-\nabla^2\rho^{(p)}(\mathbf{r}) _{bcp}$	$-\nabla^2\rho^{(u)}(\mathbf{r}) _{bcp}$	ε^d	
B_2H_6	BH_{bridge}	0.12434	0.12031	0.00403	-0.14557	-0.15582	0.01025	0.72166	
		0.13502	0.13126	0.00376	0.24530	0.24071	0.00459	0.12959	
	BH_{term}	0.18426	0.17973	0.00454	0.24203	0.22588	0.01616	0.09056	
		0.20088	0.19671	0.00417	0.64524	0.64148	0.00376	0.00977	
	$B...B^b$	0.11679	0.11371	0.00308	0.04067	0.04151	-0.00084	1.32998	
$B_2H_6^{2-}$	BH_{bridge}	0.10871	0.10512	0.00360	0.18834	0.18342	0.00492	0.11459	
	BH_{term}	0.14419	0.14012	0.00407	0.12527	0.13634	0.01107	0.67775	
	$B...B^b$								
$B_4H_{10}^c$	$B_2-H_{11(bridge)}$	0.14017	0.13659	0.00357	-0.13225	-0.14131	0.00906	0.49601	
		0.14850	0.14499	0.00351	0.31547	0.30979	0.00568	0.17307	
	$B_3-H_{11(bridge)}$	0.10342	0.10034	0.00309	-0.06856	-0.07604	0.00748	0.85837	
		0.11293	0.11034	0.00258	0.13304	0.13265	0.00039	0.37769	
	B_2-B_4	0.13846	0.13518	0.00328	0.25332	0.25203	0.00129	0.33559	
		0.13717	0.13386	0.00331	0.25375	0.25205	0.00170	0.07008	
	B_3-H_{14}	0.18500	0.18128	0.00372	0.23376	0.22043	0.01333	0.08766	
		0.20127	0.19783	0.00344	0.64837	0.64501	0.00336	0.01921	
	B_2-H_7	0.18213	0.17835	0.00377	0.17064	0.15863	0.01201	0.10678	
		0.19659	0.19304	0.00355	0.61438	0.61055	0.00382	0.08046	
	B_2-B_1	no bcp							
		0.10028	0.09786	0.00241	0.03854	0.03913	-0.00058	0.87032	
	$B_5H_9^c$	$B_2-H_{12(bridge)}$	0.12062	0.11749	0.00313	-0.06657	-0.07323	0.00666	6.45833
0.12917			0.12633	0.00284	0.21071	0.20843	0.00228	0.31604	
B_2-H_8		0.18423	0.18068	0.00354	0.20833	0.19672	0.01161	0.08693	
		0.19879	0.19547	0.33211	0.63582	0.63233	0.00349	0.09007	
B_1-H_6		0.17590	0.17242	0.00348	0.09551	0.08554	0.00997	0.00000	
		0.18970	0.18637	0.00333	0.57193	0.56802	0.00391	0.00000	
B_1-B_2		0.13859	0.13503	0.00355	0.12304	0.11660	0.00644	3.24034	
		0.13883	0.13559	0.00324	0.20903	0.20757	0.00145	0.11929	
B_2-B_3		no bcp							
		0.11741	0.11481	0.00260	0.07259	0.07327	-0.00069	0.62698	
B_2-B_4	no bcp								
	no bcp								
$B_5H_{11}^c$	$B_4-H_{16(bridge)}$	0.11840	0.11537	0.00303	-0.13818	-0.14499	0.00681	1.28377	
		0.12802	0.12520	0.00281	0.21954	0.21663	0.00291	0.27336	
	B_2-H_6	no bcp							
		no bcp							
	B_5-H_{14}	0.14511	0.14177	0.00334	-0.06305	-0.07182	0.00877	0.69875	
		0.15403	0.15083	0.00320	0.34886	0.34441	0.00446	0.19806	
	B_1-H_6	0.15054	0.14699	0.00355	-0.05909	-0.06756	0.00848	0.43287	
		0.15569	0.15217	0.00352	0.32352	0.31736	0.00616	0.24869	
	B_1-H_7	0.17958	0.17607	0.00350	0.13166	0.12080	0.01086	0.01518	
		0.19467	0.19136	0.00330	0.59993	0.59627	0.00365	0.00987	
	B_2-H_9	0.18448	0.18102	0.00346	0.23111	0.21870	0.01241	0.13271	
		0.20062	0.19742	0.00320	0.64844	0.64521	0.00323	0.03314	
	B_2-H_{11}	0.18310	0.17961	0.00350	0.20816	0.19580	0.01236	0.09965	
		0.19939	0.19614	0.00324	0.63561	0.63224	0.00337	0.02847	
	B_2-H_{14}	no bcp							
		0.11463	0.11227	0.00236	0.12920	0.12939	-0.00019	0.80530	
	B_5-H_8	0.18665	0.18312	0.00353	0.22340	0.21159	0.01181	0.07284	
0.20166		0.19836	0.00330	0.65038	0.64706	0.00332	0.06786		
B_1-B_2	no bcp								
	0.10386	0.10142	0.00244	0.09498	0.09494	0.00005	0.37470		
B_1-B_5	0.12980	0.12681	0.00299	0.19690	0.19559	0.00131	1.65077		
	0.12863	0.12563	0.00300	0.19073	0.18944	0.00129	0.18989		
B_2-B_5	0.11184	0.10905	0.00279	0.12162	0.11896	0.00266	20.37688		
	0.11309	0.11047	0.00262	0.08750	0.08689	0.00061	0.34419		
B_4-B_5	no bcp								
	0.10566	0.10325	0.00241	0.03758	0.03825	-0.00067	0.95615		

^a Second line in columns 3 to 9 for each bond indicate the densities and $L(\mathbf{r})$ at $\rho^{(u)}(\mathbf{r})$ vs(3,-1)cp. ^b There are no bond critical points for $\rho(\mathbf{r})$ between the boron atoms. ^c See Figure 1 for atoms labeling. ^d Ellipticity. ^e All quantities are in atomic units.

behavior may be noted at $\rho^{(u)}$ vs(3,-1)cp's. For the bridge bonds BH the bcp's of the total density show ρ and $\rho^{(p)}$ depletion, while $\rho^{(u)}$ concentrates at those points; but at $\rho^{(u)}$ vs(3,-1)cp all fields concentrate. The bcp of the ρ density on the bridge bond BH is closer to the B atom than the corresponding vs(3,-1)cp of $\rho^{(u)}$ density, which is located

near the midpoint between these atoms (cf. Figure 2). Hence, this density delocalizes over the bonding region, which can be interpreted as a "spill" of $\rho^{(u)}$ over the bond region, as may be observed in Figure 3(b),(d). ρ and $\rho^{(p)}$ have no bcp at the BB sequence. However a rcp appears there, as has been mentioned above, showing a large charge concentration

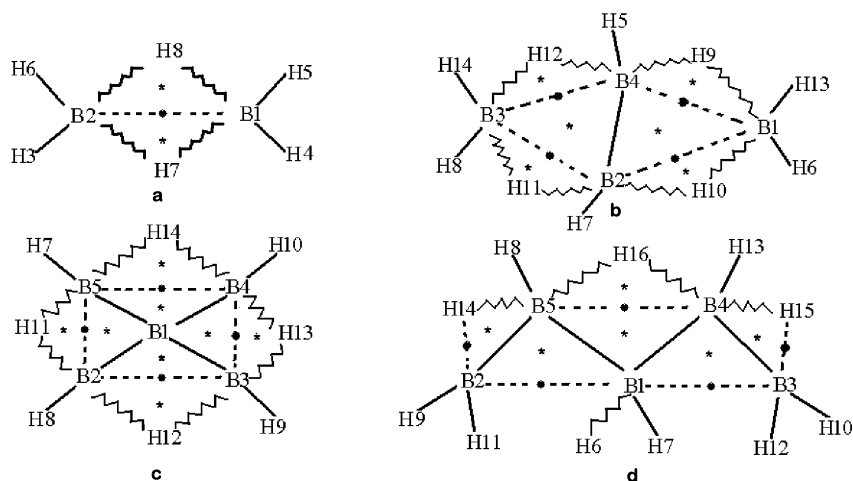


Figure 2. Schematic electronic structure of borane systems: (a) B_2H_6 , (b) B_4H_{10} , (c) B_5H_9 , and (d) B_5H_{11} . Solid lines mean all fields possess one bcp between the atoms and concentrate on it; zigzag lines mean all fields possess one bcp between the atoms, $\rho(r)$ and $\rho^{(p)}(r)$ show depletion and $\rho^{(u)}(r)$ shows accumulation; dashed lines mean only $\rho^{(u)}(r)$ vs(3, -1)cp (•) exist between the atoms and this field concentrates on it. Asterisks (*) indicate vs(3,1)cp locations.

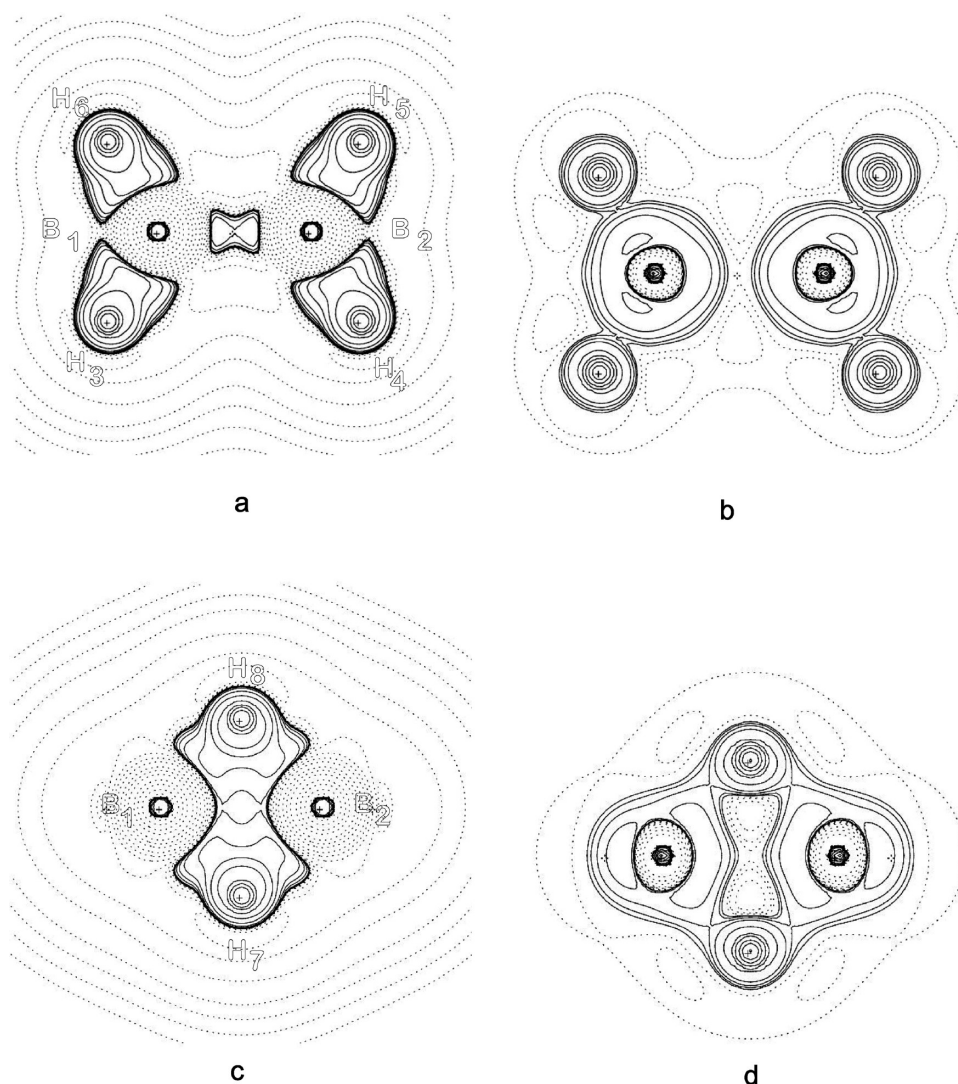


Figure 3. $L(r)$ contour maps of B_2H_6 molecule for effectively paired ((a) and (c)) and effectively unpaired densities ((b) and (d)) in the plane containing the BH_2 groups and the H_{bridge} atoms, respectively. Positive and negative values are denoted by solid and dashed lines, respectively.

(cf. Table S2 and Figure 3(a),(c)) which justifies the mentioned value of the bond order BB. Nevertheless, the

behavior of $\rho^{(u)}$ at its own vs(3, -1)cp on this sequence shows a slight concentration, as may be observed in Table 2 and

Figure 3(b),(d). The onset of the two vs(3,+1)cp's of $\rho^{(u)}$ for each bridge bond BHB disrupts the idea of a four-center bond and permits to identify the two (3c-2e) bonds, in agreement with the indicator *styx* and the results from the integrated formulation. Besides, the vs(3,-1)cp of $\rho^{(u)}$ density at the BB sequence permits to define the above expressed rule for the local formalism.

It is interesting to note the changes on the electron distribution of this system once negative charges are introduced to obtain the anion $B_2H_6^{2-}$, which is an electron nondeficient system. As may be seen in Table 1, the topologies of ρ and $\rho^{(p)}$ densities remain identical. The pairing and unpairing densities show a remarkable decrease in the bonding regions BB, and consequently a great concentration of these densities appears around the nuclei as shown by Figure S1(a)–(d). Moreover all bonds BH increase their strength, as shown by the increment of the bond orders in Table 1. The topology of $\rho^{(u)}$ does not exhibit the BB vs(3,-1)cp as in the compound B_2H_6 , and both vs(3,+1)cp's collapse into only one. Therefore, the above established rule seems to be violated by the increments of the two-center bond order values, and consequently the population of the three-center bridge BHB notably decreases. The local view of the broken down (3c-2e) structures in this system may be noted from the contour maps by the lack of concentration zone between boron atoms (Figure S1(a),(c)) and the missing of vs(3,-1)cp in the line joining the mentioned atoms. It may also be noted that the onset of $L(\mathbf{r})$ concentration of both ρ and $\rho^{(p)}$ fields (cf. Tables 2 and S2) at the bridge bonds BH make them to look like typical terminal bonds BH.

B_4H_{10} . The geometrical structure of this system is shown in Figure 1, and its characterization is given by the *styx* number 4012 (see Table 1). Let us analyze each of the structural indicators in view of the hypothesis established for the previous system. $s = 4$ means that this compound has four bridge three-center bonds BHB. The values of three-center bond index Δ_{BHB} shown in Table 1 correspond to four bonds formed by the atoms B(1)H(10)B(2) and B(1)H(9)B(4) and their symmetrical ones, with bond index values of 0.086 and 0.087, respectively. These values are similar to those found in the B_2H_6 system. The bridge two-center bond orders I_{BH} forming the (3c-2e) bonds have intermediate values in comparison with terminal bonds BH in boranes although one of them, the bond order for the atoms B(2)H(10) and its symmetric B(4)H(12), B(2)H(11), and B(4)H(9) have a value slightly smaller than the normal one for a typical bond BH (cf. the discussion for B_2H_6 system). Thus the nonlocal rule applied to diborane is confirmed for the bonds BHB in this system. Therefore, the structural $s = 4$ number and the nonlocal results are coincident.

The local information contained in ρ (and $\rho^{(p)}$) shows ncp's over each of the nuclei and bcp's for each terminal bond BH (B(1)H(6), B(1)H(13), B(2)H(7) and their symmetric ones B(3)H(8), B(3)H(14), B(4)H(5)), for each bridge bond BH and for the B(2)B(4) bond, indicating that the atoms of each mentioned pair are bonded. Besides, an rcp for the total density appears for the B(1)H(10)B(2)B(4)H(9) atomic sequence (see Table 1). As has been discussed for the B_2H_6

system, it may be interpreted as a five-center bond, hiding the existence of (3c-2e) bonds predicted by the integrated formalism and the structural *styx* numbers for this structure. However, the information contained in $\rho^{(u)}$ shows the onset of two vs(3,-1)cp's indicating a subtle interaction between the atomic pairs B(1), B(2) and B(1), B(4) (and a similar situation is observed for their symmetric ones B(3), B(2) and B(3), B(4)). This find as well as the onset of two vs(3,+1)cp's just defining the sequences B(1)H(9)B(4) and B(1)H(10)B(2), and other two for their symmetrical BHB sequences, seem to indicate four bonds BHB. Consequently there is one vs(3,-1)cp for each pair of atoms forming the three-center sequences, and besides there is also a vs(3,+1)cp for each of the mentioned sequences. Hence the local version of the integrated rule for detecting three-center bonds holds, and it is also coincident with both structural and nonlocal descriptors regarding BHB bonds.

The $t = 0$ number of *styx* indicates that this system has no bonds BBB. The normal BB bond order value for the pair B(2)B(4) (0.476) and the small ones for the B(1)B(2) pair and the remaining BB pairs in the system (0.095) break the rule for the three-center bond nonlocal description and is in agreement with a very small (3c-2e) bond B(1)B(2)B(4) of 0.056. Therefore, the structural and nonlocal descriptions are coincident and do not predict these bonds for the system. The local description of ρ and $\rho^{(p)}$ shows a bcp for the B(2)B(4) bond but not for B(1)B(2) and B(1)B(4) ones, i.e., the local view shows that there is a bcp only for one of the pairs of atoms in the BBB sequence. However the topology of $\rho^{(u)}$ shows a vs(3,-1)cp related with each of these pairs, and we also found a vs(3,+1)cp for the sequence BBB, insinuating a subtle BBB interaction. This information contained in the $\rho^{(u)}$ topology is not available from the description provided by ρ and $\rho^{(p)}$ densities and may be interpreted as an actual but not too strong interaction between three boron atoms which is not detected by the other descriptions, although is concordant with a very small (3c-2e) population for the sequence BBB. This result would induce to recognize a very weak (3c-2e) bond for both the B(1)B(2)B(4) sequence and its symmetrical B(2)B(3)B(4) and confirms that the information in $\rho^{(u)}$ goes beyond the structural description by the other methodologies. Figure 4(a),(b) corresponds to maps for $L(\mathbf{r})$ in the plane containing the boron atoms (1), (2), and (4), while Figure 4(c),(d) refers to the plane defined by the sequence B(1)H(10)B(2). These maps show the concentration of effectively paired density and specially the unpaired one in the bonding regions. The "spillage" of the last density seems to be of crucial importance to form three-center bonds in electron deficient systems, as indicated by the existence of vs(3,-1)cp's only in $\rho^{(u)}$ density.

The value of the y number predicts only one BB bond. As explained above, the nonlocal description is in agreement with this prediction for the atoms B(2) and B(4). For the sake of completeness we incorporate Figure 4(e),(f) in the plane defined by atoms (1), (2), and (11), which permit to note the pairing density concentration, showing graphically this weak BB interaction; however, as explained above this interaction does not stand for a true BB bond and may be

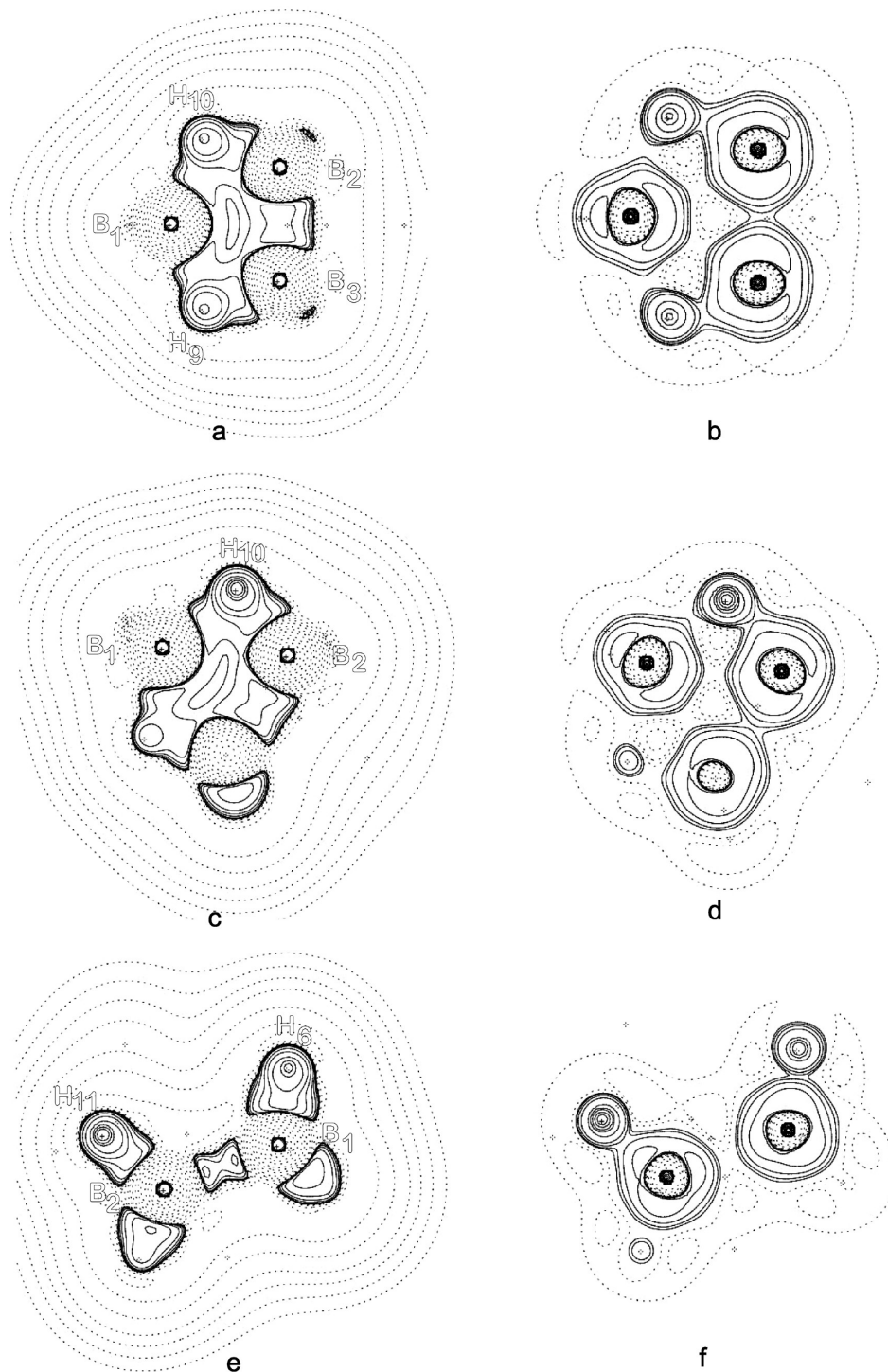


Figure 4. $L(\mathbf{r})$ contour maps of B_4H_{10} molecule for effectively paired ((a), (c), and (e)) and effectively unpaired densities ((b), (d), and (f)) in the planes defined by atoms 1-2-4, 1-10-2, and 1-2-11, respectively (Figure 1). Positive and negative values are denoted by solid and dashed lines, respectively.

called a *quasi-bond* because the atoms are involved in BHB bonds. The last number, $x = 2$, shows that for BH_2 groups all descriptions are coincident. Similarly to the diborane case, the three density fields show electron concentration at bcp's for the bonds BH of each group BH_2 and for the bonds B(2)H(7) and B(4)H(5). For the bridge bonds BH, ρ and $\rho^{(p)}$ show depletion at the bcp's while $\rho^{(u)}$ concentrates; however, all fields concentrate at $\rho^{(u)}$ vs(3,-1)cp's. The "spill" of $\rho^{(u)}$ over the bonding region is again observed at

all bonds (cf. parts (b), (d), and (f) of Figure 4), thus showing positive values of $L(\mathbf{r})$ (of $\rho^{(u)}$) at all vs(3,-1)cp's (see Table 2). This $\rho^{(u)}$ behavior was not found in none of the classical bonds analyzed in previous papers^{32,33} and therefore seems to be typical of electron deficient systems.

B_5H_9 . As illustrated in Figure 1, the hydride B_5H_9 belongs to the C_{4v} symmetry group and is structurally characterized by the *styx* number 4120 (see Table 1). It means that this compound has four bridge three-center bonds BHB ($s = 4$),

only one bond BBB ($t = 1$), two bonds BB ($y = 2$), and no groups BH₂ ($x = 0$). The sequence B(2)H(12)B(3) and its other three symmetrical ones (cf. Figure 2) have BH bond orders of 0.364, as shown in Table 1. These are intermediate values but not normal values for such a bond (see e.g. the diborane), thus it fulfills the nonlocal rule to form a (3c-2e) bond. Hence, the nonlocal description is coincident with the structural one. From the local point of view, each bond BH in the sequences BHB has (3, -1) cp's for all the densities (bcp for total density), but we only found a vs(3, -1)cp of $\rho^{(u)}$ between the boron atoms. Accordingly, a $\rho^{(u)}$ vs(3, +1)cp just defining the sequence B(2)H(12)B(3) is found, which with its symmetrical ones is indicating four bonds BHB. Again, the unpaired density subtly connects these atoms resulting in a (3c-2e) bond and the local rule also holds. In conclusion, the three descriptions are coincident. The $t = 1$ and $y = 2$ numbers for this system indicate only a bond BBB and two two-center bonds BB. These results are not compatible with the symmetry group exhibited by the system. However, to rescue such structural indicators, four equivalent resonance structures have been invoked to fit them to quantum concepts.^{34,37,43} The nonlocal analysis shows bond orders values of about (0.4) between the axial boron and each equatorial boron atom, which indicates an interaction. The same happens between each pair of equatorial borons, but their bond order values are lower than the formers (of about 0.3). Thus, there are intermediate bond orders between each pair formed by one equatorial boron and the axial boron and consequently four (3c-2e) BBB bonds are present. The nonlocal picture is reinforced from the local point of view by the information of the unpaired electron density, which shows four vs(3, -1)cp's on the equatorial bonds BB and the corresponding four vs(3, +1)cp's, one for each equatorial/axial/equatorial sequence BBB. These results show the usefulness of the local and nonlocal descriptions and a drawback for *styx* numbers, which may not account for the symmetries in a direct manner, i.e., local and nonlocal pictures are able to overcome the ambiguities found in the localization schemes. Each boron atom is also bonded to a hydrogen one by a normal bond (see in Table 1 $I_{B(1)H(6)} = 0.627$ and the four bond indices BH (equatorial boron) with the value 0.601). Hence, the structural description in terms of x number is coincident with the nonlocal point of view.

In Table 2 it may be noted that at each bcp of the total density for bonds BH of the equatorial boron atoms (hydrogen atoms numbers (7), (8), (9), and (10)), all density fields show concentration and a small ellipticity of 0.08693, which confirms the existence of normal BH bonds as discussed above; a similar behavior of these fields is observed at the vs(3, -1)cp of $\rho^{(u)}$. Nevertheless, the bonds equatorial/bridge BH show different behavior than the previous ones; the pairing density field depletes while the unpaired one concentrates at the bcp and an opposite behavior is found for that field at the vs(3, -1)cp of $\rho^{(u)}$ density. This result is similar to that found in the diborane and once more indicates the "spillage" of the unpaired density over the bonding regions. The graphical representation of these results may be observed on the $L(\mathbf{r})$ maps of both densities, which are shown in Figure 5(a),(b) on the plane containing the boron

atoms (1), (2), and (3) and in Figure 5(c),(d) on the plane defined by equatorial boron atoms. The analysis of the bonds BB axial/equatorial type (B(1)B(2) in Table 2) permits to observe that at the total density bcp's all fields concentrate, and consequently these bonds are similar to the BB bonds found in the system B₄H₁₀. However, the values of $L(\mathbf{r})$ for the axial-equatorial boron in this system decrease and the ellipticity values are largely increased for the pairing density, but the $L(\mathbf{r})$ values for the unpaired density are largely increased, in comparison with the sequence B(2)B(4) in B₄H₁₀. These marked differences may be interpreted as the sharing electron cloud by the atoms involved in the set of four "resonance/fractional bonds",^{34,37,43} in the local analysis framework.

B₅H₁₁. The geometrical information and the structure representation of B₅H₁₁ are shown in Figures 1 and 2, respectively. Table 1 reports the value 3203 for its *styx* number. It means that this system possesses three three-center bonds BHB, two bonds BBB, no two-center bonds BB, and three groups BH₂. As may be seen in Figure 2, the candidate sequences to form bridge bonds BHB are B(4)H(16)B(5) and B(2)H(14)B(5) (and their symmetrical B(3)H(15)B(4)). The system exhibits only one rcp formed by the sequence B(5)B(1)B(4)H(16) which, as in the previous systems, may not clearly discriminate the (3c-2e) bonds formed by the atoms involved in it. Table S2 shows the behavior of the densities at the rcp. Once again the appearance of the vs(3, +1)cp permits to establish the correct three-center bonds as shown in Table 1. According to this table the sequence B(4)H(16)B(5) shows a three-center population of 0.105, bond orders BH of 0.348 (each of them characterized by a bcp for the total density and a vs(3, -1)cp for the $\rho^{(u)}$ one), and a bond order BB of 0.172 with only a vs(3, -1)cp of $\rho^{(u)}$ density (cf. Table 1) indicating any information not available from the total density. The BH bonding regions show similar features to those found in the diborane molecule: depletion of the pairing density while the unpaired density concentrate in the bcp and an opposite behavior for $\rho^{(p)}$ at the vs(3, -1)cp (cf. Table 2). The density at the bcp is lower than in the other analyzed similar sequences, indicating its weakness, and it also shows a very high ellipticity (Table 2). Similar features are found for the other two above-mentioned bridge sequences BHB. The pair of bondings B(4)H(15) and B(4)H(13) deserve some comments; the first one possesses the topological features which characterize a typical bridge bond BH, i.e., pairing density depletion and unpaired density concentration at the bcp and concentration of the three fields at vs(3, -1)cp. The other one sequence, B(4)H(13), is a typical terminal bond BH as in the diborane, and, thus, the local analysis permits to establish that these two BH bonds do not conform a group BH₂. Graphically, these considerations may be observed from Figure 6(a),(b) and Figure 6(c),(d) showing $\rho^{(p)}$ and $\rho^{(u)}$ $L(\mathbf{r})$ maps in the planes of the atoms (4), (5), and (16) and in the plane of the atoms (4), (5), and (15), respectively. Moreover we found a vs(3, -1)cp between each of the pairs B(4)B(3) (B(2)B(5)), B(2)H(14) (B(3)H(15)), and B(5)H(14) (B(4)H(15)), and a (3, +1)cp for the B(3)B(4)H(15) (B(2)B(5)H(14)) sequences.

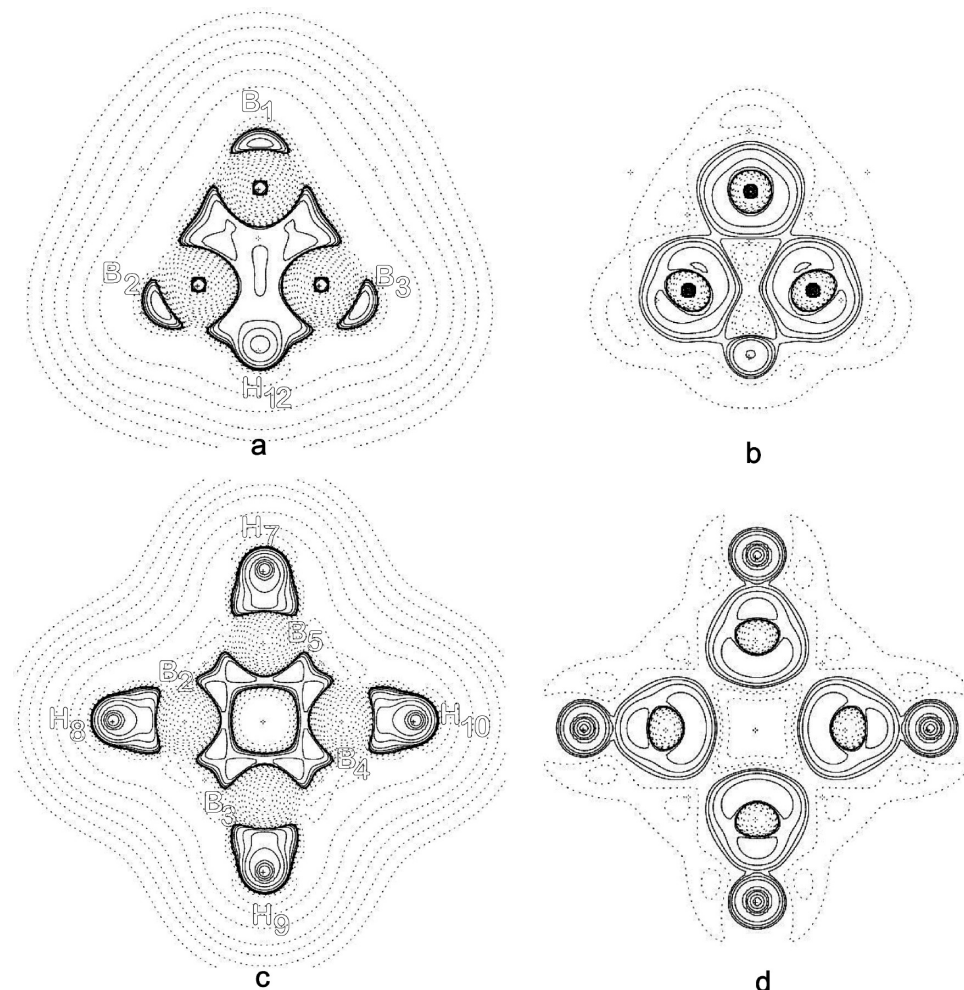


Figure 5. $L(\mathbf{r})$ contour maps of B_5H_9 molecule for effectively paired ((a) and (c)) and effectively unpaired densities ((b) and (d)) in the plane defined by atoms 1-2-3 and 2-3-4-5, respectively (Figure 1). Positive and negative values are denoted by solid and dashed lines, respectively.

Therefore, there are three bridge bonds BHB in this system, and, thus, the structural, local, and nonlocal descriptions coincide.

Let us inspect the BBB sequences B(1)B(4)B(5) and B(1)B(2)B(5) (and its symmetrical B(1)B(3)B(4)). Table 1 shows a three-center population of 0.113 for the sequence B(1)B(4)B(5); the bonds BB involved in this three-center bonding have values of bond order of 0.459 for B(1)B(4) (B(1)B(5)) which are very close to normal values for such a bond (quasi-normal bond) and 0.172 for B(4)B(5) (weak BB bond). The local analysis shows similar features for the two other mentioned BBB series B(1)B(2)B(5) and B(1)B(3)B(4) as reported in Table 1. However, the bond orders for the BB pairs involved in these three-center bonds are lower (see e.g. 0.214 for B(1)B(2) in Table 1). Hence, the fulfillment of the rule of existence of fractional two-center bond orders to identify three-center bondings is better in these cases. It permits to conclude that B(1)B(4)B(5) might be considered as a weak three-center bond thus justifying the value $t = 2$. Figure S2(a),(b), in the plane of B(1), B(3), B(4), and Figure S2(c),(d), in the plane of B(1), B(4), B(5), show the accumulation of the densities in the space between those atoms. Regarding the BB bonds, there are four bonds [B(1)B(4) and B(1)B(5) (quasi-normal BB bonds) and

B(2)B(5) and B(3)B(4) (weak BB bonds)] all of them exhibiting a bcp and a $vs(3,-1)cp$. These results indicate at first glance four BB bonds. Nevertheless, as all these BB pairs are involved in three-center BBB bondings, they may not be considered as normal two-center bonds, in agreement with the value $y = 0$ of the *styx* number. It is worthy to note the behavior of the densities for the BB pairs, as shown in Tables 1 and 2. The atom B(4) appears to be bonded to the atoms B(1) and B(3). These sequences show small concentration for the three density fields at bcp and at $vs(3,-1)cp$ and exhibit a great ellipticity which indicates an instability interaction between the boron atoms. The interaction B(4)B(1) shows concentration of the three fields at both bcp and $vs(3,-1)cp$. Both interactions exhibit low density at these points, indicating a weak character interaction, high ellipticity for the first sequence, and extremely high for the later one, which points out instability; consequently both of them may not be considered as normal bonds from the local point of view. Figure S2(a),(b) shows the density $L(\mathbf{r})$ maps. For the other B(4)B(5), B(1)B(2), and B(1)B(3) interactions we only found a $vs(3,-1)cp$, which means that a true bond between them may not be considered. This scenario induces the formation of three-center bonds BBB, in agreement with the above interpretation of the integrated description. As has been

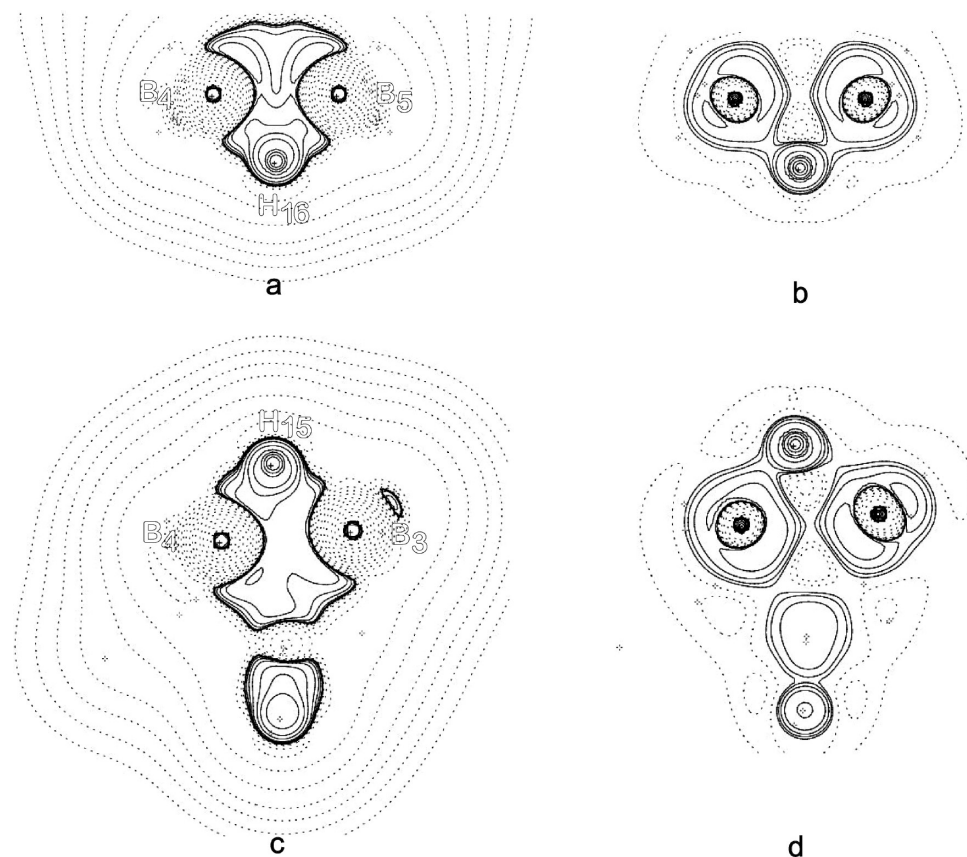


Figure 6. $L(\mathbf{r})$ contour maps of B_5H_{11} molecule for effectively paired ((a) and (c)) and effectively unpaired densities ((b) and (d)) in the plane defined by atoms 5-4-16 and 3-4-15, respectively (Figure 1). Positive and negative values are denoted by solid and dashed lines, respectively.

indicated for the system B_4H_{10} , in which we have used the term *quasi-bonds* for such weak interactions, the concentration of the pairing component between $B(4)B(5)$ and $B(1)B(2)$ ($B(1)B(3)$) may clearly be noted in Figure S3(a),(c) in the planes defined by atoms $B(4)B(5)H(8)$ and $B(1)B(2)B(3)$, respectively; Figure S3(b),(d) is the corresponding one to the unpaired density. Concerning the BH_2 groups, previous arguments lead us to consider only the sequences formed by $B(2)H(9)H(11)$, its symmetric $B(3)H(10)H(12)$ and $B(1)H(6)H(7)$ as candidates to conform such moieties. The bond orders and the bond critical points for the total density associated with each bond BH , as shown in Table 1, indicate that there are three groups BH_2 . Hence, the local and nonlocal descriptions are in agreement with the value $x = 3$ in the structural number *styx*.

3.2. General Discussion. As a general feature, it may be noted (Table 2) that in the bridge bonds BH the $L(\mathbf{r})$ of the pairing density depletes, while the unpaired one concentrates at the bcp. Concerning the BB and terminal BH bonds, both of them show Laplacian values of identical sign at the bcp's for the $\rho^{(p)}$ and $\rho^{(u)}$ density fields. It was observed in a previous work³³ that this behavior usually indicates an important ionic component of the bond. However, in the present case, the electronic distribution is different since this trend remains at the $vs(3,-1)cp$'s. Exception to this feature may be observed at the sequence $B(1)H(6)$ in the system B_5H_{11} ; in this case the Laplacian of the pairing density depletes at the bcp while that of $\rho^{(u)}(\mathbf{r})$ concentrates. At $vs(3,-1)cp$ both fields concentrate. This behavior is typical

of the bridge bonds BH , like $B(3)H(11)$ in B_4H_{10} or $B(2)H(12)$ in B_5H_9 one. However, the observation is concordant with the appreciable bond order value between the atoms $H(6)$ and $B(2)$ in this B_5H_{11} system, despite there are not critical points for any field in such sequence, i.e., it does not constitute a bond. Consequently, the small but appreciable three-center population $B(1)H(6)B(2)$ (see Table 1) can be explained through the population $B(1)H(6)$ and the $vs(3,-1)cp$ existing between each pair of atoms in this sequence. These observations indicate a very special "subtle" interaction between all these atoms. The Laplacian density maps for these bonds show a spillage of the unpaired density. Table S1 shows that the pairing density at bridge atoms H , i.e., the hydrogen atoms involved in the $(3c-2e)$ bonds, exhibits the greatest concentration in comparison with the terminal hydrogen ones. This can be interpreted as a bigger accumulation of the density on those atoms, even for the total density. Hence, it induces the bridge BH bcp localizes further from the hydrogen atom than in terminal sequences BH . It is reinforced by the information in Table 2 and in the contour maps of $L(\mathbf{r})$ for $\rho^{(p)}(\mathbf{r})$, showing that in the valence region there is depletion close to the boron atom while concentration appears close to the hydrogen atoms, while for $L(\mathbf{r})$ of $\rho^{(u)}(\mathbf{r})$ it is noted concentration all over bonding region. Conventional covalent and ionic bonds are characterized by concentration of the paired and depletion of the unpaired density Laplacian and depletion of both of them, respectively.³³ The present case of bondings does not exhibit this behavior, but it is featured by the concentration of the unpaired field at all

cp's; the paired component shows concentration at terminal BH and BB bonds and depletion at bridge bonds BH; this is the most important feature of such kind of bonding which physically means that this density is induced toward the interatomic regions, as may be clearly observed in the contour maps.

Regarding the (3c-2e) bonds BHB, in the topological local description the pairing density accumulation between boron atoms in the bridge is relevant. An inspection of all the systems studied reveals the concentration of the pairing component while it is observed depletion for the unpaired one at the vs(3,-1)cp in that kind of bonds. Although no bcp is found, such a situation is typical of a covalent bond. Therefore, it may be considered as subtle information contained in the unpaired density which allows one to infer the existence of a *quasi-bond* between the boron atoms in agreement with the appreciable two- and three-center populations of these atomic sequences. An important relation arises from the data observation of Tables 2 and S1. The Laplacian of the density $\rho^{(u)}(\mathbf{r})$ presents concentration at the bcp of bridge bonds BH and at the ncp of the bridge atom H; while the integrated populations of the sequences bridge BHB and BB increase, the concentration of that density field at the mentioned points depletes. Another interesting feature to analyze these systems is to introduce the concepts of competitive and cooperative phenomena, due to the fact that boron atoms are involved in different three-center bonds. In Table 1 it may be noted the joint growing tendency of the BHB population and that of the corresponding BB one in the compound series B_2H_6 , B_4H_{10} , B_5H_{11} , and B_5H_9 with values 0.061, 0.086, 0.105, and 0.118 for three-center bondings and 0.056, 0.095, 0.172, and 0.385 for two-center ones, respectively. To explain this fact, we will choose the system B_4H_{10} as a reference in which the population BB related with each one of the four bonds BHB is not shared by any other interaction BHB, and consequently there is a relation one to one (1:1) for each type of interaction. In diborane the two (3c-2e) bonds BHB share the BB population, and both populations diminish from the reference value of the compound B_4H_{10} . A possible interpretation of these situations is to think that the BHB interactions compete to distribute the electrons within the structure with a (2:1) relation. This idea is supported by the analysis of the remaining systems in which the trends noted above are fulfilled, i.e., BHB and BB jointly grow and the increment of BB population is favored if the boron atoms are involved in BBB bonds instead of involved in the BHB ones. Therefore, it seems that BHB bonds are competitive between them and cooperative with BBB bonds.

4. Final Remarks and Conclusions

In this work we have studied electron distributions in some electron deficient molecules (boranes) which are prototype systems possessing no conventional patterns of bonding. The studied systems have been chosen to cover the main features for this type of molecular structures. We have attempted to extend to these systems our methodology, developed and applied in previous works to describe successfully more classical bondings. The challenge is to describe these

nonclassical systems under two approaches: the topological nonlocal (integrated) procedure and the local one, showing the complementary information arising from both treatments. We have compared the results with those provided by the structural indices or *styx* numbers, universally accepted to establish the structure of the boron hydrides. The results reported in this work show that both topological formalisms are in agreement with the structural one and go beyond it in cases where the electronic description becomes ambiguous, pointing out the subtle information contained in the density $\rho^{(u)}(\mathbf{r})$, which may not be obtained from the $\rho(\mathbf{r})$ and $\rho^{(p)}(\mathbf{r})$ ones. These results allow one to establish the local rule, which has been formulated within the discussion of the B_2H_6 system, in analogy with the nonlocal one for detecting the atomic sequences forming a (3c-2e) bond beyond the Lewis model.

The main difference found between the electron distributions of these systems and those arising from the classical ones is that the paired part of the density remains localized in bonding and nuclear regions while the unpaired one is not only near the nuclear positions but spills out over the bonding regions, as trying to compensate the electron deficiency. Our results prove that the effectively unpaired density, $\rho^{(u)}(\mathbf{r})$, provides important information which may not be directly observed from the total and paired densities, constituting a suitable tool to measure the deviations from Lewis bonding models in these compounds.

Acknowledgment. This work has been partly supported by the Projects X-017 (Universidad de Buenos Aires), PIP No. 5098/05 (Consejo Nacional de Investigaciones Científicas y Técnicas, República Argentina), the Spanish Ministry of Education (Grant No. CTQ2006-01849/BQU), and the Universidad del País Vasco (Grant No. GIU06/03). We thank the Universidad del País Vasco for allocation of computational resources. R.M.L. acknowledges aid from the Universidad de la Cuenca del Plata (Corrientes, Argentina) for facilities provided during the course of this work.

Supporting Information Available: Density and $L(\mathbf{r})$ of total density ρ components: $\rho^{(p)}$, $\rho^{(u)}$ at the nuclear critical points of total density at the CISD/6-31G** level of calculation. All quantities are in atomic units (Table S1); density and $L(\mathbf{r})$ for the total density ρ and its contributions $\rho^{(p)}$, $\rho^{(u)}$ at ring critical points of total density at the CISD/6-31G** level of calculation^a. All quantities are in atomic units (Table S2). $L(\mathbf{r})$ contour maps of the $B_2H_6^{2-}$ molecular anion for effectively paired ((a) and (c)) and effectively unpaired densities ((b) and (d)) in the plane containing the BH_2 groups and the H_{bridge} atoms, respectively. Positive and negative values are denoted by solid and dashed lines, respectively (Figure S1); $L(\mathbf{r})$ contour maps of the B_5H_{11} molecule for effectively paired ((a) and (c)) and effectively unpaired densities ((b) and (d)) in the plane defined by atoms 1-3-4 and 1-4-5, respectively (Figure 1). Positive and negative values are denoted by solid and dashed lines, respectively (Figure S2); $L(\mathbf{r})$ contour maps of the B_5H_{11} molecule for effectively paired ((a) and (c)) and effectively unpaired densities ((b) and (d)) in the plane defined by atoms 4-5-8 and 1-2-3, respectively (Figure 1). Positive and

negative values are denoted by solid and dashed lines, respectively (Figure S3). This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Giambiagi, M.; Giambiagi, M. S.; Grepel, D. R.; Heymann, C. D. *J. Chim. Phys.* **1975**, *72*, 15.
- (2) Mayer, I. *Chem. Phys. Lett.* **1983**, *97*, 270.
- (3) Mayer, I. *Int. J. Quantum Chem.* **1986**, *29*, 73.
- (4) Bochicchio, R. C. *THEOCHEM* **1991**, 228, 209, and references therein.
- (5) Bochicchio, R. C.; Lain, L.; Torre, A. *Chem. Phys. Lett.* **2003**, *374*, 576, and references therein.
- (6) Alcoba, D. R.; Bochicchio, R. C.; Lain, L.; Torre, A. *Chem. Phys. Lett.* **2007**, *442*, 157.
- (7) Bochicchio, R. C.; Lain, L.; Torre, A. *Chem. Phys. Lett.* **2003**, *375*, 45.
- (8) Bochicchio, R. C. *THEOCHEM* **1998**, 429, 229.
- (9) Lain, L.; Torre, A.; Bochicchio, R. C.; Ponec, R. *Chem. Phys. Lett.* **2001**, *346*, 283.
- (10) Takatsuka, K.; Fueno, T.; Yamaguchi, K. *Theor. Chim. Acta* **1978**, *48*, 175.
- (11) Takatsuka, K.; T. Fueno, T. *J. Chem. Phys.* **1978**, *69*, 661.
- (12) Staroverov, V. N.; Davidson, E. R. *Chem. Phys. Lett.* **2000**, *330*, 161.
- (13) McWeeny, R. *Methods of Molecular Quantum Mechanics*; Academic: London, 1969; see also references therein.
- (14) Davidson, E. R. *Reduced Density Matrices in Quantum Chemistry*; Academic: New York, 1976; see also references therein.
- (15) Bamzai, A. S.; Deb, B. M. *Rev. Mod. Phys.* **1981**, *53*, 95.
- (16) Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure*; Macmillan Publ. Co.: New York, 1982.
- (17) Mulliken, M. S. *J. Chem. Phys.* **1955**, *23*, 1833.
- (18) Robby, K. *Mol. Phys.* **1974**, *27*, 81.
- (19) Wiberg, K. *Tetrahedron* **1968**, *24*, 1083.
- (20) Salvador, P.; Mayer, I. *J. Chem. Phys.* **2004**, *120*, 5046.
- (21) Hirshfeld, F. L. *Theor. Chim. Acta* **1977**, *44*, 129.
- (22) Davidson, E. R.; Chakravorty, S. *Theor. Chim. Acta* **1992**, *44*, 129.
- (23) Torre, A.; Alcoba, D. R.; Lain, L.; Bochicchio, R. C. *J. Phys. Chem. A* **2005**, *109*, 6587.
- (24) Clark, A. E.; Davidson, E. R. *Int. J. Quantum Chem.* **2003**, *93*, 384.
- (25) Bader, R. F. W. *Atoms in Molecules: A Quantum Theory*; Clarendon Press: Oxford, U.K., 1994; see also references therein.
- (26) Popelier, P. L. A. *Atoms in Molecules: An Introduction*; Pearson Edu.: London, 1999.
- (27) Fradera, X.; Austen, M. A.; Bader, R. F. W. *J. Phys. Chem. A* **1999**, *103*, 304.
- (28) Torre, A.; Lain, L.; Bochicchio, R. *J. Phys. Chem. A* **2003**, *107*, 127.
- (29) Lain, L.; Torre, A.; Bochicchio, R. *J. Phys. Chem. A* **2004**, *108*, 4132.
- (30) Cioslowski, J.; Mixon, S. T. *J. Am. Chem. Soc.* **1991**, *113*, 4142.
- (31) Bader, R. F. W. *Chem.--Eur. J.* **2006**, *12*, 7769.
- (32) Lobayan, R. M.; Bochicchio, R. C.; Lain, L.; Torre, A. *J. Chem. Phys.* **2005**, *123*, 144116.
- (33) Lobayan, R. M.; Bochicchio, R. C.; Lain, L.; Torre, A. *J. Phys. Chem. A* **2007**, *111*, 3166.
- (34) Wade, K. *Electron Deficient Compounds. Studies in Modern Chemistry*; T. Nelson and Sons. Ltd.: London, 1971.
- (35) Fox, M. A.; Wade, K. *Pure Appl. Chem.* **2003**, *75*, 1315, and references therein.
- (36) Wade, K. *Nat. Chem.* **2009**, *1*, 92.
- (37) Eriks, K.; Lipscomb, W. N.; Schaeffer, R. *J. Chem. Phys.* **1954**, *22*, 754.
- (38) Coleman A. J.; Yukalov V. I. *Reduced Density Matrices: Coulson's Challenge; Lecture Notes in Chemistry*; Springer: Berlin, 2000; Vol. 72.
- (39) Popelier, P. L. A. *Coord. Chem. Rev.* **2000**, *197*, 169.
- (40) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, revision C.02*; Gaussian, Inc.: Wallingford, CT, 2004.
- (41) Biegler-Köning, F. W.; Bader, R. F. W.; Tang, T. H. *J. Comput. Chem.* **1982**, *13*, 317.
- (42) Gillespie, R. J.; Popelier, P. L. A. *Chemical Bonding and Geometry*; Oxford U. P.: New York, 2001.
- (43) Bochicchio, R.; Ponec, R.; Uhlik, P. *Inorg. Chem.* **1997**, *36*, 5363, and references therein.

CT900281Q

JCTC

Journal of Chemical Theory and Computation

Unexpected Trimerization of Pyrazine in the Coordination Sphere of Low-Valent Titanocene Fragments

Thomas Jung,[†] Rüdiger Beckhaus,^{*,†} Thorsten Klüner,^{*,‡} Sebastian Höfener,[§] and Wim Klopper^{*,§}

Inorganic Chemistry and Theoretical Chemistry, Institute of Pure and Applied Chemistry, Carl von Ossietzky University, D-26111 Oldenburg, Germany and Institute of Physical Chemistry, Universität Karlsruhe (TH), D-76128 Karlsruhe, Germany

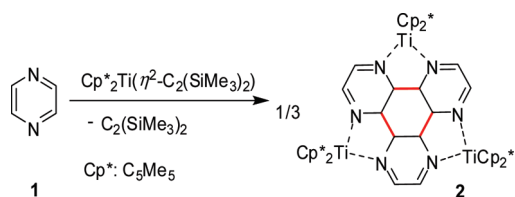
Received March 23, 2009

Abstract: The titanium mediated trimerization of pyrazine leads to the formation of a tris–chelate complex employing a 4a,4b,8a,8b,12a,12b-hexahydrodiypyrazino[2,3-*f*:2',3'-*h*]quinoxaline ligand (HATH₆, **3**). The driving force in the formation of the (Cp*₂Ti)₃(HATH₆) complex **2** is attributed to the formation of six Ti–N bonds. We show that density functional theory (DFT) fails to predict quantitatively correct results. Therefore, post-Hartree–Fock methods, such as second-order Møller–Plesset perturbation theory (MP2), in combination with coupled-cluster (CC) methods must be used. Both MP2 and CCSD(T) levels of theory provide endothermic trimerization energies, showing that the plain pyrazine trimer is not stable with respect to decomposition into its monomers. Complete basis set (CBS) results for the MP2 level of theory were computed using explicitly correlated wave functions. With these, we estimate the CCSD(T) CBS limit of the hypothetical trimerization energy to be +0.78 eV. Thus, the trimerization is facilitated by the formation of six Ti–N bonds with a calculated formation energy of –1.32 eV per bond.

Introduction

One of the general advantages of organometallic compounds is their ability to stabilize highly reactive, short-lived organic molecules and fragments in the coordination sphere of transition metals. Some remarkable examples are arynes and aryne complexes,¹ vinylidenes and vinylidene complexes,² and cyclobutadiene and its complexes.³ In the course of our investigations concerning the self-assembled formation of multinuclear highly ordered complexes, employing low-valent early transition-metal complexes and aromatic bis-azines as bridging ligands, different reaction patterns have been found.⁴ In these investigations, the formation of well-

Scheme 1. Formation of (Cp*₂Ti)₃(HATH₆) (**2**)



defined molecular squares or rectangles,⁵ dehydrogenative^{6,7} as well as C–C coupling reactions induced by the primary formation of radical anions have been observed (Scheme 1).⁸

The formation of the trinuclear titanium complex **2** by a spontaneous 3-fold C–C coupling of three pyrazine molecules (**1**) inspired us to investigate theoretically the hypothetical C–C coupling reaction of three pyrazine molecules (Scheme 2), forming the 4a,4b,8a,8b,12a,12b-hexahydrodiypyrazino[2,3-*f*:2',3'-*h*]quinoxaline ligand (HATH₆, **3**).

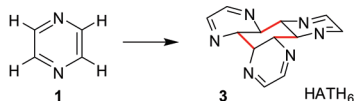
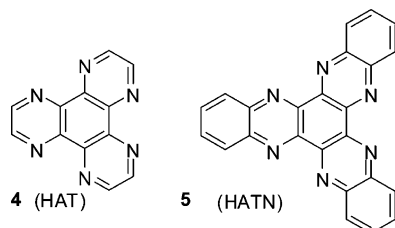
To the best of our knowledge, the nonaromatic molecule HATH₆ (**3**) is still unknown, in contrast to the well-known aromatic hexaazatriphenylene ligand (HAT) (**4**). However,

* Corresponding authors. Telephone: +49 441 798 3656 (R.B.); +49 441 798 3681 (T.K.); and +49 721 608 7263 (W.K.). Fax: +49 441 798 3851 (R.B.); +49 441 798 3964 (T.K.); and +49 721 608 3319 (W.K.). E-mail: ruediger.beckhaus@uni-oldenburg.de (R.B.); thorsten.kluener@uni-oldenburg.de (T.K.); and klopper@chem-bio.uni-karlsruhe.de (W.K.).

[†] Inorganic Chemistry, Carl von Ossietzky University.

[‡] Theoretical Chemistry, Carl von Ossietzky University.

[§] Universität Karlsruhe (TH).

Scheme 2. Formation of HATH₆ (**3**)**Chart 1.**

the chemistry employing **4** is often limited by the low solubility and the laborious synthesis.⁹ HAT derivatives are generally interesting ligands for different applications. The tris-chelating nitrogen heterocyclic molecule 1,4,5,8,9,12-hexaazatriphenylene (HAT) was often studied in the context of its coordination modes to metal ions,¹⁰ photophysical properties,¹¹ liquid crystalline ordering,¹² light-harvesting functionality,¹³ chirality,¹⁴ and DNA related chemistry.¹⁵ When employing quinoxaline instead of pyrazine in the reaction **1**→**2**, the formation of 1,6,7,12,13,18-hexaazatrinaphthylene (HATN, **5**) via the corresponding [(Cp₂Ti)₃(HATN)] complex is observed.⁶ Under oxidative conditions, **5** can be obtained in quantitative yields. In attempts to use this oxidative workup procedure for **2**, no hints of **3** are found. Only traces of **4** are obtained (Chart 1).

The molecular symmetry of HAT derivatives leads to especially interesting building blocks creating self-assembled 3D frameworks.¹⁶ Employing paramagnetic metal centers in

HAT-type complexes, magnetic properties of those derivatives,¹⁷ including spin-frustrated behavior or antiferromagnetic couplings, are of interest.¹⁸

Under experimental conditions, **2** is characterized by the formation of an ideal cyclohexane ring in the chair conformation, exhibiting all hydrogens in axial positions. By comparing the experimental⁸ structural parameters of the titanium complex **2** with the hypothetical molecule **3** (studied theoretically at the MP2/cc-pVTZ level), it is found that the enamide form **A** becomes dominant for **2**, whereas the diazadiene structure is preferred in **3** (Figure 1).

The dominating enamide form **A** for **2** becomes obvious by comparing the C–C and C–N distances in **2** and **3** (Figure 1).⁸

The Csp³–Csp³ bond lengths of the central cyclohexane rings of **2** and **3** are nearly identical (**2**: av. 1.53 Å; **3**: 1.52 Å). The Csp²–Csp² bond lengths of **2** and **3** show slight differences (**2**: av. 1.41 Å; **3**: 1.47 Å). In the same way, the Csp²–N distances are longer in **2** (av. 1.345 Å) than in **3** (1.29 Å).

Theoretical Study

To shed light on the energetics of the pyrazine trimerization, we calculated the corresponding reaction energy at various levels of theory. The results we obtained first are summarized in Table 1, where we report reaction energies for two different basis sets applying density functional theory (DFT) and ab initio quantum chemical methods. Regardless of the level of theory, we found the pyrazine trimerization to be energetically unfavorable with reaction energies in the range of 0.64 to 2.32 eV. Although this result is not unexpected as such, the large range of energies is quite disturbing. Analyzing our results in detail, we found that DFT generally

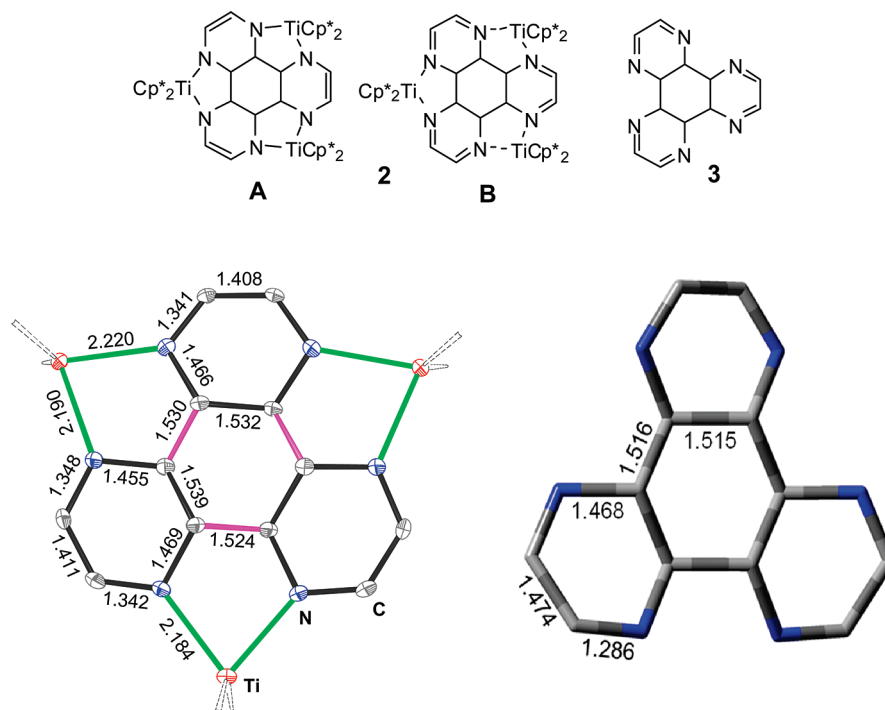


Figure 1. C–C and C–N distances (Å) in the titanium complex **2** (X-ray structure, left) and the hypothetical molecule **3** (right) (Ti red, N blue, C gray).

Table 1. Endothermic Trimerization Energy (in eV) As Obtained at Various Levels of Theory (Full Geometry Optimization)^a

basis	method	ΔE /eV
cc-pVDZ	B3LYP ^{26–29}	1.98
	PBE ³⁰	1.43
	PW91 ^{31,32}	1.40
	PBE1W ³³	1.77
	MPWLYP1W ³³	2.15
	HF	1.90
	MP2	1.04
	CCSD//MP2	0.64
	CCSD(T)//MP2	0.68
	cc-pVTZ	B3LYP ^{26–29}
PBE ³⁰		1.70
PW91 ^{31,32}		1.70
PBE1W ³³		2.07
MPWLYP1W ³³		2.55
MPWB1K ^{34–36}		1.42
HF		2.01
MP2		1.11
CCSD//MP2		0.71

^aThe CCSD and CCSD(T) results were obtained at the corresponding MP2/cc-pVDZ and MP2/cc-pVTZ geometries (denoted as //MP2). Only valence orbitals were correlated.

fails to predict accurate results for this endothermic reaction, underestimating the stability of the pyrazine trimer by up to 1.5 eV.^{19–22} Note that the very popular B3LYP hybrid exchange–correlation functional seems to give the worst results of any mean-field method, including Hartree–Fock (HF). Recent functionals such as PBE1W and MPWLYP1W do not improve the situation, although the MPWB1K functional turns out to be superior to other exchange–correlation functionals. Taking our CCSD and CCSD(T) calculations as a reference, MP2 gives much more accurate energies than that of HF or DFT, although an error of about 0.5 eV remains. To further analyze this phenomenon and to estimate the MP2 reaction energy at the complete basis set (CBS) limit, we subsequently performed MP2-F12 calculations in which we found an endothermic trimerization energy of 1.14 eV (Table 2).

The trimerization of pyrazine is endothermic by about 0.64–0.68 eV at the coupled-cluster level in a small double- ζ basis (CCSD/cc-pVDZ and CCSD(T)/cc-pVDZ values). In this same basis, the MP2 value amounts to 1.04 eV, and hence, our best estimate for the trimerization energy is 0.78 eV. This value is expected to represent the CCSD(T) basis set limit. In view of the large variation in the DFT results, we conclude that an accurate treatment of electron correlation is crucial to obtain quantitatively correct results. Here, accuracy refers to both excitation level (up to connected triples) and one-particle basis set (F12 approach).

For a realistic treatment of the whole reaction, the pyrazine trimerization was furthermore studied in the presence of three equivalents of TiCl₂, leading to **6**, which serves as a model compound for **2** in which the Cp* ligand is substituted by chlorides. This choice is motivated by computational savings and validated by previous experimental studies.²³

In view of the large computational effort, we restricted these calculations to MP2 studies using a cc-pVDZ basis set. Although the error in the trimerization energy is at least 0.3–0.4 eV, as suggested by the results reported in Tables 1 and 2, the qualitative interpretation of our results is straightforward. Due to the formation of six Ti–N bonds, the reaction of **1**→**6** becomes strongly exothermic, exhibiting a reaction energy of about –8.70 eV, assuming a high-spin configuration of maximum multiplicity. This results in a formation energy of about –1.32 eV (127 kJ mol^{–1}) per Ti–N bond (calculated from (8.70 – 0.78 eV)/6), which is a reasonable value in good agreement with rare experimental values for titanium(II) complexes employing π -acceptor ligands, as shown for Cp₂Ti(η^2 -PhN=NPh) (144 ± 12 kJ mol^{–1}).²⁴ In the case of titanium(IV) amido derivatives, higher mean bond enthalpies are found (e.g., 334 ± 10 kJ mol^{–1} for Cp₂Ti(NC₈H₆)₂).²⁵ Thus, despite the fact that the pyrazine trimerization is energetically unfavorable, we predict a strong stabilization of the trinuclear titanium complex **2**.

Hence, our quantum chemical calculations reveal how Ti-based coordination chemistry can be used to stabilize a pyrazine trimer species, which corresponds to a highly endothermic reaction in a gas phase. Further experimental and theoretical studies will explore this new potential of possible reactivity.

Conclusions

The formation of the pyrazine trimerization product in form of the trinuclear titanium complex **2** illustrates in a spectacular manner the advantages of organometallic synthesis. Due to the formation of six moderately strong Ti–N bonds, the pyrazine trimer is stabilized, although a trimerization of pyrazine as such is an energetically unfavorable process, that is, an endothermic reaction. As our large-scale high-quality ab initio calculations reveal, such a trimerization is energetically clearly unfavorable, exhibiting a hypothetical formation energy of +0.78 eV. It should be noted that the pyrazine trimerization reaction turns out to be a prominent example of a complete failure of popular DFT functionals. Along these lines, PBE and PW91 results are virtually identical and strongly overestimate the endothermic reaction energy by almost 1 eV. The widely used hybrid functional B3LYP exhibits an even larger error of about 1.5 eV, which is worse than Hartree–Fock. The best DFT result has been obtained using the MPWB1K functional still resulting in an error of about 0.7 eV with respect to our coupled-cluster calculations. Thus, the pyrazine trimerization reaction turns out to be a prominent example in which an accurate correlation treatment is mandatory. Our large-scale correlation calculations, which explore the limits of currently available computing resources, allow for a systematic and highly accurate calculation of the pyrazine trimerization energy. Eventually, we demonstrate

Table 2. Endothermic Trimerization Energy (in eV) as Obtained at the Frozen-core MP2-F12 Level^a

basis	CABS	MP2 fitting basis	JK fitting basis	Hartree–Fock	MP2	MP2-F12
def2-TZVP	def2-SVP	def2-QZVPP	def2-QZVPP	2.30	2.11	1.33
def2-TZVP	def2-TZVP	def2-QZVPP	def2-QZVPP	2.30	2.11	1.31
aug-cc-pVTZ	cc-pVTZ	aug-cc-pVQZ	aug-cc-pVTZ	2.15	0.94	1.06
aug-cc-pV5Z	cc-pVTZ	aug-cc-pV5Z	aug-cc-pV5Z	2.19	1.13	1.14

^a Using Ansatz 2B and a Slater-type geminal exponent of $\gamma = 1.4 a_0^{-1}$. The frozen-core MP2/cc-pVTZ equilibrium structures were used for all calculations. An MP2 fitting basis was used as complementary auxiliary basis set (CABS).

Table 3. MP2 Valence Shell Correlation Energies (in Hartree) for the Benzene Molecule with Various Basis Sets^d

basis	basis size	ΔE_{MP2}	% ^a	$\Delta E_{\text{MP2-F12 fixed}}^b$	% ^a	$\Delta E_{\text{MP2-F12 optimized}}^c$	% ^a
aug-cc-pVTZ	414	-0.9645	91.2	-1.0505	99.3	-1.0534	99.6
aug-cc-pV5Z	1242	-1.0358	97.9	-1.0571	100.0	-1.0573	100.0
aug-cc-pV6Z	1896	-1.0446	98.8	-1.0576	100.0	-1.0576	100.0
def2-TZVP	222	-0.9325	88.2	-1.0432	98.6	-1.0483	99.1
def2-QZVPP	522	-1.0081	95.3	-1.0544	99.7	-1.0555	99.8

^a The percentage with respect to the estimated MP2 basis set limit. ^b The F12 amplitudes were held fixed according to the rational generator approach. ^c The F12 amplitudes were optimized variationally. ^d The limit is estimated to be $-1.0576 \pm 0.0005 E_h$.⁵⁸

in a quantitative manner how the formation of a pyrazine trimer as a trinuclear titanium complex can be rationalized.

Experimental Section

Synthesis. The trinuclear complex **2** was prepared by reaction of $\text{Cp}^*_2\text{Ti}(\eta^2\text{-C}_2(\text{SiMe}_3)_2)$ with pyrazine in a 1:1 ratio, as described earlier.⁸

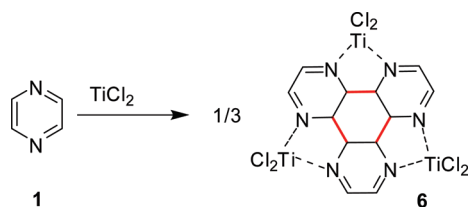
Pyrazine Trimerization 1→3. Almost all calculations concerning the energetics of the pyrazine trimerization **1**→**3** were performed using the Gaussian03³⁷ quantum chemistry program package. Only the explicitly correlated MP2-F12 calculations were performed with an experimental version of the Turbomole v5.10 code.³⁸ Using correlation-consistent basis sets of cc-pVDZ and cc-pVTZ-quality,³⁹ we applied different DFT approaches, including the B3LYP, PBE, PW91, PBE1W, MPWLYP1W, and MPWB1K functionals for exchange and correlation. For a more accurate treatment of electron correlation, we performed conventional ab initio calculations at the MP2 level. The energy of the trimerization reaction was calculated performing full geometry optimizations of reactants and products for each basis set and level of theory. In order to improve our prediction of the energetics further, we performed CCSD and CCSD(T) calculations using the MP2-optimized structures for each basis set.

Explicitly Correlated Calculations. To explore errors due to truncation of the one-particle basis set, we performed MP2-F12 calculations^{40,41} in the basis sets def2-TZVP,⁴² aug-cc-pVTZ,^{39,43} and aug-cc-pV5Z.^{39,43} For the trimer, the latter basis comprises as many as 3 246 contracted basis functions, and, apart from the purpose of calculating accurate second-order trimerization energies, the MP2-F12/aug-cc-pV5Z calculations were also performed to assess the performance of our newly developed code.^{44,45} The MP2-F12 calculations were performed using ansatz 2B⁴⁰ with a complementary auxiliary basis set (CABS) to approximate three electron integrals.⁴⁶ Two electron integrals were approximated using robust density fitting.⁴⁷ A Slater-type geminal⁴⁸ with exponent $\gamma = 1.4 a_0^{-1}$ was used as explicitly correlated two-particle basis function (represented by a six-term Gaussian expansion⁴⁹). To circumvent two-electron

integrals involving commutators of kinetic energy and Fock operators, we applied an approach similar to that of Kedžuch and co-workers.⁵⁰ The F12 amplitudes were held fixed, following the rational generator approach of Ten-no.⁵¹ The MP2-F12 model has been implemented in the module RICC2 of Turbomole developed by Hättig and co-workers.^{52,53} As CABS, we used the def2-SVP, def2-TZVP, and cc-pVTZ cbas sets of Weigend et al.^{54,55} that usually are used for Coulomb fitting in RI-MP2 theory (MP2 fitting basis). For the density fitting approach itself within both the MP2 and F12 parts, we used the def2-QZVPP and aug-cc-pVXZ ($X = Q, 5$) cbas sets of Hättig.⁵⁶ To represent the Fock operator within the MP2-F12 calculation of the correlation energy, we used the JK fitting basis sets (jkbasis) of Weigend.⁵⁷

To be able to judge the quality of the explicitly correlated calculations, we have added results for benzene of which the MP2 complete basis set limit is well-known (Table 3).⁵⁸ These calculations were carried out using the Slater exponent $\gamma = 1.5 a_0^{-1}$ and the Coulomb fitting basis def2-QZVPP as CABS. The geometry was taken from ref 58. Table 3 shows that 99.95% of the valence shell correlation energy of benzene is recovered at the MP2-F12/aug-cc-pV5Z level using fixed amplitudes, and we expect that the MP2-F12/aug-cc-pV5Z calculations of the reaction energy of the trimerization of pyrazine are equally accurate.

As usual, the convergence of the conventional results is rather slow, even in the augmented sextuple-zeta basis with more than 1 800 contracted basis functions, less than 99% of the MP2 correlation energy limit is obtained. Table 3 demonstrates clearly a significantly improved convergence behavior when explicit correlation is taken into account in the wave function. Even using fixed amplitudes, the MP2-F12 calculation in a small def2-TZVP basis with only 222 contracted basis functions is almost as accurate as the standard MP2 calculation in an aug-cc-pV6Z basis. When the F12 amplitudes are optimized variationally, the small basis even outperforms this conventional result. Nevertheless, the differences among fixed and variationally optimized F12 amplitudes are small. Hence, for larger systems such as the pyrazine trimer, the fixed amplitudes approach is favorable because fixing the amplitudes results in much lower com-

Scheme 3. Formation of **6**

putational costs. Note that in the aug-cc-pV5Z basis used for the pyrazine trimer, 100.0% of the correlation energy of benzene is obtained with fixed as well as optimized amplitudes.

Pyrazine Trimerization 1→6. All calculations for the trimerization reaction 1→6 were performed at the MP2 level of theory with full geometry optimization employing for Ti a Karlsruhe^{40,41} DZ basis set augmented by one polarization set of f-type functions with an exponent of $0.8 a_0^{-2}$. All remaining atoms were represented using a cc-pVDZ set. In all calculations of Ti-containing compounds, maximum-spin multiplicity was assumed, that is, TiCl_2 was calculated in a triplet state, which could be verified by subsequent calculations, while the complex **6** was treated as a septet.

Acknowledgment. The development of the MP2-F12 method used in the present work has been supported by the Deutsche Forschungsgemeinschaft (DFG) through the Priority Program SPP 1145 (KI 721/2-3). It has been further supported by the DFG through SPP 1118 (Be 1400/4-1). S.H. thanks the Deutsche Telekom Stiftung for a PhD studentship stipend.

References

- Buchwald, S. L.; Nielsen, R. B. *Chem. Rev.* **1988**, *88*, 1047–1058.
- Bruce, M. I. *Chem. Rev.* **1998**, *98*, 2797–2858.
- Seyferth, D. *Organometallics* **2003**, *22*, 2–20.
- Beckhaus, R. In *Activating Unreactive Substrates: The Role of Secondary Interactions/Bis-azines in the coordination sphere of early transition metals*; Bolm, C., Hahn, F. E., Eds.; WILEY-VCH: Weinheim, Germany, 2009; p 183–207.
- Kraft, S.; Hanuschek, E.; Beckhaus, R.; Haase, D.; Saak, W. *Chem.—Eur. J.* **2005**, *11*, 969–978.
- Piglosiewicz, I. M.; Beckhaus, R.; Saak, W.; Haase, D. *J. Am. Chem. Soc.* **2005**, *127*, 14190–14191.
- Piglosiewicz, I. M.; Beckhaus, R.; Wittstock, G.; Saak, W.; Haase, D. *Inorg. Chem.* **2007**, *46*, 7610–7620.
- Kraft, S.; Beckhaus, R.; Haase, D.; Saak, W. *Angew. Chem., Int. Ed.* **2004**, *116*, 1609–1614.
- Rademacher, J. T.; Kanakarajan, K.; Czarnik, A. W. *Synthesis* **1994**, 378–380.
- Browne, W. R.; Hage, R.; Vos, J. G. *Coord. Chem. Rev.* **2006**, *250*, 1653–1668.
- Ishi-I, T.; Yaguma, K.; Kuwahara, R.; Taguri, Y.; Mataka, S. *Org. Lett.* **2006**, *8*, 585–588.
- Ishi-I, T.; Hirayama, T.; Murakami, K. I.; Tashiro, H.; Thiemann, T.; Kubo, K.; Mori, A.; Yamasaki, S.; Akao, T.; Tsuboyama, A.; Mukaide, T.; Ueno, K.; Mataka, S. *Langmuir* **2005**, *21*, 1261–1268.
- Ishi-I, T.; Murakami, K. I.; Imai, Y.; Mataka, S. *J. Org. Chem.* **2006**, *71*, 5752–5760.
- Rutherford, T. J.; Keene, F. R. *J. Chem. Soc., Dalton Trans.* **1998**, 1155–1162.
- Smith, J. A.; Morgan, J. L.; Turley, A. G.; Collins, J. G.; Keene, F. R. *Dalton Trans.* **2006**, 3179–3187.
- Baxter, P. N. W.; Lehn, J. M.; Baum, G.; Fenske, D. *Chem.—Eur. J.* **1999**, *5*, 102–112.
- Marshall, S. R.; Rheingold, A. L.; Dawe, L. N.; Shum, W. W.; Kitamura, C.; Miller, J. S. *Inorg. Chem.* **2002**, *41*, 3599–3601.
- Shatruck, M.; Chouai, A.; Prosvirin, A. V.; Dunbar, K. R. *Dalton Trans.* **2005**, 1897–1902.
- Furche, F.; Perdew, J. P. *J. Chem. Phys.* **2006**, *124*, 044103.
- Woodcock, H. L.; Schaefer, H. F., III; Schreiner, P. R. *J. Phys. Chem A* **2002**, *106*, 11923–11931.
- Patey, M. D.; Dessent, C. E. H. *J. Phys. Chem. A* **2002**, *106*, 4623–4631.
- Zhao, Y.; Truhlar, D. G. *Acc. Chem. Res.* **2008**, *41*, 157–167.
- Böhme, U.; Beckhaus, R. *J. Organomet. Chem.* **1999**, 585, 179–188.
- Dias, A. R.; Dias, P. B.; Diogo, H. P.; Galavao, A. M.; de Piedade, M. E. M.; Simoes, J. A. M. *Organometallics* **1987**, *6*, 1427–1432.
- Calhorda, M. J.; da Costa, R. G.; Dias, A. R.; Simões, J. A. M. *Dalton Trans.* **1982**, 2327–2330.
- Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785–789.
- Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, *98*, 11623–11627.
- Vosko, S. H.; Wilk, L.; Nusair, M. *Can. J. Phys.* **1980**, *58*, 1200–1211.
- Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
- Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1992**, *46*, 6671–6687.
- Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1993**, *48*, 4978.
- Dahlke, E. E.; Truhlar, D. G. *J. Phys. Chem. B* **2005**, *109*, 15677–15683.
- Adamo, C.; Barone, V. *J. Chem. Phys.* **1998**, *108*, 664–675.
- Becke, A. D. *J. Chem. Phys.* **1996**, *104*, 1040–1046.
- Zhao, Y.; Truhlar, D. G. *J. Phys. Chem. A* **2004**, *108*, 6908–6918.
- Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.;

- Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. Gaussian, Inc.: Wallingford, CT, 2004.
- (38) *Turbomole*, version 5.10; Turbomole GmbH: Karlsruhe, Germany, 2008. See: www.turbomole.com.
- (39) Dunning, T. H., Jr. *J. Chem. Phys.* **1989**, *90*, 1007–1023.
- (40) Klopper, W.; Manby, F. R.; Ten-no, S.; Valeev, E. F. *Int. Rev. Phys. Chem.* **2006**, *25*, 427–468.
- (41) Werner, H.-J.; Adler, T. B.; Manby, F. R. *J. Chem. Phys.* **2007**, *126*, 164102.
- (42) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297–3305.
- (43) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796–6806.
- (44) Bischoff, F. A.; Höfener, S.; Glöb, A.; Klopper, W. *Theor. Chem. Acc.* **2008**, *121*, 11–19.
- (45) Höfener, S.; Bischoff, F. A.; Glöb, A.; Klopper, W. *Phys. Chem. Chem. Phys.* **2008**, *10*, 3390–3399.
- (46) Valeev, E. F. *Chem. Phys. Lett.* **2004**, *395*, 190–195.
- (47) Manby, F. R. *J. Chem. Phys.* **2003**, *119*, 4607–4613.
- (48) Ten-no, S. *Chem. Phys. Lett.* **2004**, *398*, 56–61.
- (49) Tew, D. P.; Klopper, W. *J. Chem. Phys.* **2005**, *123*, 074101.
- (50) Kedžuch, S.; Milko, M.; Noga, J. *Int. J. Quantum Chem.* **2005**, *105*, 929–936.
- (51) Ten-no, S. *J. Chem. Phys.* **2004**, *121*, 117–129.
- (52) Hättig, C.; Weigend, F. *J. Chem. Phys.* **2000**, *113*, 5154–5161.
- (53) Hättig, C.; Köhn, A. *J. Chem. Phys.* **2002**, *117*, 6939–6951.
- (54) Weigend, F.; Haser, M.; Patzelt, H.; Ahlrichs, R. *Chem. Phys. Lett.* **1998**, *294*, 143–152.
- (55) Weigend, F.; Köhn, A.; Hättig, C. *J. Chem. Phys.* **2002**, *116*, 3175–3183.
- (56) Hättig, C. *Phys. Chem. Chem. Phys.* **2005**, *7*, 59–66.
- (57) Weigend, F. *J. Comput. Chem.* **2007**, *29*, 167–175.
- (58) Yamaki, D.; Koch, H.; Ten-no, S. *J. Chem. Phys.* **2007**, *127*, 144104.

CT900137R

Reaction Path Optimization with Holonomic Constraints and Kinetic Energy Potentials

Jason B. Brokaw,[†] Kevin R. Haas,[‡] and Jih-Wei Chu^{*‡}

*Department of Chemistry and Department of Chemical Engineering,
University of California, Berkeley, California 94720*

Received March 24, 2009

Abstract: Two methods are developed to enhance the stability, efficiency, and robustness of reaction path optimization using a chain of replicas. First, distances between replicas are kept equal during path optimization via holonomic constraints. Finding a reaction path is, thus, transformed into a constrained optimization problem. This approach avoids force projections for finding minimum energy paths (MEPs), and fast-converging schemes such as quasi-Newton methods can be readily applied. Second, we define a new objective function – the total Hamiltonian – for reaction path optimization, by combining the kinetic energy potential of each replica with its potential energy function. Minimizing the total Hamiltonian of a chain determines a minimum Hamiltonian path (MHP). If the distances between replicas are kept equal and a consistent force constant is used, then the kinetic energy potentials of all replicas have the same value. The MHP in this case is the most probable isokinetic path. Our results indicate that low-temperature kinetic energy potentials (<5 K) can be used to prevent the development of kinks during path optimization and can significantly reduce the required steps of minimization by 2–3 times without causing noticeable differences between a MHP and MEP. These methods are applied to three test cases, the C_{7eq}-to-C_{ax} isomerization of an alanine dipeptide, the ⁴C₁-to-¹C₄ transition of an α-D-glucopyranose, and the helix-to-sheet transition of a GNNQQNY heptapeptide. By applying the methods developed in this work, convergence of reaction path optimization can be achieved for these complex transitions, involving full atomic details and a large number of replicas (>100). For the case of helix-to-sheet transition, we identify pathways whose energy barriers are consistent with experimental measurements. Further, we develop a method based on the work energy theorem to quantify the accuracy of reaction paths and to determine whether the atoms used to define a path are enough to provide quantitative estimation of energy barriers.

Introduction

Characterizing rare events in molecular systems is of critical importance in many fields of chemistry, material science, and biology. A large amount of effort has, thus, been devoted to developing computational methods to assist in identifying transition states and reaction paths.^{1–21} For transitions between two metastable states, a simple and powerful

strategy is to employ a chain of replicas of the molecular system to connect one state to another.^{2,3} A reaction path is obtained by optimizing an objective function of the chain. For example, minimizing the total potential energy of a chain determines a minimum energy path (MEP).⁴ Other commonly employed objective functions for path optimization will be discussed later. Many computational methods such as elastic band,^{2,3} nudged elastic band (NEB),^{5–9} max-flux path,^{10,11} action-derived molecular dynamics,¹² string,^{13–19} harmonic Fourier bead,²⁰ and replica path²¹ are based on the chain-of-states framework. Significant progress has been

* Corresponding author. E-mail: jwchu@berkeley.edu.

[†] Department of Chemistry.

[‡] Department of Chemical Engineering.

made in different fields by applying these methods to understand the mechanisms of important transition processes.^{22,23}

The major challenge of chain-of-states methods is ensuring chain continuity while optimizing the specified objective function, and this nature of dual objectives severely limits the stability, efficiency, and robustness of reaction path optimization.⁵⁻²¹ Problems typically encountered include the overestimation of transition barriers and conditional convergence of reaction path optimization, especially when studying complex transitions. These issues also make it difficult to quantify the accuracy of reaction path optimization. In this work, we develop systematic approaches to overcome the difficulties resulting from the dual objectives of chain-of-states methods to broaden the applications of the chain-of-states methods. Two methodologies are developed: (1) use holonomic constraints to ensure equal distances between replicas and (2) use kinetic energy potentials to prevent the development of kinks during path optimization. Furthermore, we develop a method to quantify the accuracy of a reaction path by employing the work energy theorem. Our results indicate that the developments of this work significantly enhance the stability, efficiency, and robustness of reaction path optimization using a chain of replicas and enable applications to complex molecular systems. In the following, difficulties due to the dual objectives of chain-of-states methods will first be summarized. We then describe the two strategies for overcoming these difficulties and our approach of quantifying the accuracy of reaction path optimization.

To ensure chain continuity, ways to control the spacing between replicas along a chain are required.^{2,3} Without loss of generality, we focus on maintaining equal distances between replicas. A commonly employed strategy is to add potentials that restrain the distances between replicas to the same value.^{2,3} However, optimizing the combined objective function results in solutions that satisfy neither the equal-spacing requirements nor the criteria of optimizing an objective function.^{5,6} In cases of optimizing MEPs, gradients of potential energy functions prevent restraint potentials from equally spacing replicas and cause sliding-down; gradients of restraint potentials prevent replicas from reaching minima on the hyperplanes perpendicular to a path and cause corner-cutting.^{5,6}

To decouple the effects of the objective function and restraint potentials, a nudged elastic band (NEB) method was invented to project away force components that cause sliding-down and corner-cutting based on an estimation of tangent vectors.^{5,6} Although a path with stationary NEB forces is indeed a MEP, projected forces are not conservative.^{5,8} As a result, it is difficult to apply faster converging quasi-Newton methods for path optimization.^{8,24} A formal strategy is to separate the dual objectives of equal spacing and optimizing the objective function when applying quasi-Newton methods,⁸ but this approach makes the underlying optimization problem underdetermined. An alternative strategy is to employ the magnitude of projected forces as the objective function,^{25,26} but this requires the Hessian matrices of replicas for computing a gradient vector and is, thus, limited to relatively small molecular systems.

Similar to NEB, the zero-temperature string (ZTS) method employs projected forces for finding a MEP.¹⁵ Recently, a simplified string method was proposed that does not require explicit force projections.¹⁷ Unprojected forces, i.e., directions of steepest descent, are used to update the configurations of replicas; sliding-down, caused by force components parallel to the path, is then removed by reparametrization.¹⁷ Reparametrization is based on a continuous representation of a curve from a finite number of points via numerical interpolation and is a major difference between ZTS^{15,17} and NEB;^{5,6} in NEB, restraint potentials are used to control distances between replicas. As such, ZTS and the simplified string method both exhibit linear convergence as a steepest descent optimization.¹⁷

In summary, the dual objectives of path optimization, i.e., finding an optimal solution of the specified objective function and ensuring chain continuity, make it difficult to apply fast-converging quasi-Newton schemes for path optimization. This difficulty has limited the stability, efficiency, and robustness of chain-of-state methods, especially when a large number of degrees of freedom and a rugged potential energy surface are involved. To overcome these challenges, a plausible strategy is to enforce the equal-spacing requirements as holonomic constraints during path optimization. In this way, analytical theory can be precisely implemented in the form of constrained optimization without imposing ad hoc procedures such as adding restraint potentials or reparametrization. As a result, force projections are no longer required, and the Lagrange multipliers are used to balance the constraint forces and true forces with high numerical precision. With equal distance holonomic constraints, quasi-Newton methods and other optimization schemes that assume conservative forces can, thus, be readily applied. Parameterization of a continuous curve from a finite number of replicas also becomes unnecessary, thus, eliminating the dependence of path optimization on interpolation schemes.^{13,14} Using holonomic constraints to control the distances between replicas also allows a straightforward implementation of general definitions of distances such as via a noncommutative rms (root-mean-squared) best-fit procedure,²⁷ which is particularly important for simulating transition processes of macromolecules.^{8,21} In this work, we devise a simple, fast, and stable scheme to equally space replicas along a chain as constraints.

In addition to maintaining equal distances, ways to prevent the development of kinks (high curvatures) are also important for stable path optimization.¹³⁻¹⁹ The tendency of developing kinks during path optimization increases with the number of replicas in a chain as well as the magnitude of forces involved in a transition.⁶ For methods that rely on force projections, this intrinsic instability also depends on the ways tangent vectors are estimated and can be improved by adopting a definition based on the relative energies between replicas.⁶ Another cause of highly curved paths is the underlying ruggedness of the potential energy (or free energy) surface. Ways of controlling curvatures along a path are important for stable and efficient path optimization. Commonly employed approaches include using angles or self-avoiding potentials or applying smoothing procedures.

Preventing kinks during path optimization introduces additional ad hoc components that may bias the results; therefore, physics based approaches are highly desired. We propose a new strategy of using potentials that represent the kinetic energy of each replica to prevent the development of kinks during path optimization. Kinetic energy potentials have a functional form that is quadratic in the mass-weighted distance between two replicas with a force constant, corresponding to the inverse of squared time. Combining kinetic energy potentials with the potential energy functions of replicas defines the total Hamiltonian of a chain, and minimizing this objective function determines a minimum Hamiltonian path (MHP). To our knowledge, this is the first time that a Hamiltonian objective function and MHP have been introduced. Different ways of setting the force constants of kinetic energy potentials can be devised. In this work, we consider a simple strategy of using a constant value for all kinetic energy potentials. By keeping the same distance between replicas, this approach ensures a constant kinetic energy along a chain and gives an isokinetic path. A fixed value of kinetic energy can also be maintained during path optimization by rescaling the force constants. A convenient way for estimating the magnitude of a kinetic energy potential is through the value of the corresponding temperature. With zero kinetic energy potentials, the objective function becomes the total potential energy, and a MEP is, thus, a 0 K MHP. In the high-temperature limit, kinetic energy potentials ensure that a straight line connecting two structures is the result of path optimization. Therefore, employing kinetic energy potentials provides a physically based approach to prevent the development of kinks during path optimization. For optimizing MEPs using low-temperature kinetic energy potentials, our results indicate that the convergence of path optimization is accelerated by 2–3 times.

In the rest of this paper, the details of equal distance holonomic constraints and kinetic energy potentials are presented. To illustrate the enhanced stability, efficiency, and robustness of path optimization, methods developed in this work are applied to find the reaction paths of three molecular transitions, the C_{7eq}-to-C_{ax} isomerization of an alanine dipeptide, the ⁴C₁-to-¹C₄ transition of an α-D-glucopyranose, and the helix-to-sheet transition of a GNNQQNY heptapeptide. Finally, the conclusion is stated.

Methodology

A. Distance between Replicas. For a molecular system composed of N atoms, a configuration state is specified by the three-dimensional position vectors, r_i ($i = 1, N$), of each atom. A general way for defining the distance (ΔI^J) between two states (replicas) I and J is^{8,21}

$$\Delta I^J = \sqrt{\frac{\sum_{i=1}^N w_i (r_i^I - \mathbf{U}^{IJ} r_i^J)^2}{\sum_{i=1}^N w_i}} \quad (1)$$

In eq 1, w_i is the weighting factor associated with each atom for measuring distance. If w_i corresponds to the mass of atom

i , eq 1 determines the mass-weighted distance between configurations I and J. For atoms that are not directly involved in the process of interest, such as solvent molecules far away from the active site of an enzyme, w_i may be set to zero.^{8,21} When computing distances between molecular structures, a common practice is to exclude the contribution from rigid body rotation and translation. The contribution from rigid body translation can be removed by putting the weighted center at the origin; to remove the contribution from rigid body rotation, we adopt an rms best-fit procedure to calculate a three-by-three unitary rotation matrix (\mathbf{U}^{IJ} in eq 1) that minimizes the distance between I and J using a Lagrangian.²⁷ The rms best-fit procedure is widely used for measuring structural differences between biomolecules; it has also been employed in the replica path method for reaction path optimization.^{8,21}

B. The Chain-of-States Framework and Equal Distance Constraints. Given two metastable states of a molecular system of N atoms with their configurations denoted as $\mathbf{r}^0 = \{r_i^0, i = 1, N\}$ and $\mathbf{r}^K = \{r_i^K, i = 1, N\}$, a chain of $K + 1$ replicas can be constructed to connect the two states. To ensure chain continuity and resolution of path, we propose to constrain the distances between replicas being equal during path optimization, i.e.,

$$\Delta I^0 = \dots = \Delta I^l = \dots = \Delta I^{K-1} = \overline{\Delta l} \quad (2)$$

Here, ΔI^l is a shorthand notation for $\Delta I^{l,l+1}$ in eq 1, and $\overline{\Delta l}$ is the averaged distance between replicas along a chain and is computed on the fly during path optimization. With $K + 1$ replicas in a chain, eq 2 involves a set of K coupled algebraic equations that need to be solved at each step of path optimization. Following the procedures of maintaining holonomic constraints in molecular simulations,²⁸ we propose the following scheme for solving eq 2 with the two ends, \mathbf{r}^0 and \mathbf{r}^K , fixed at the current position:

(i) Calculate the averaged distance, $\overline{\Delta l}$, between replicas from $\{\mathbf{r}^{(0)(0)} \dots \mathbf{r}^{(K)(0)}\}$. The superscript “(n)” specifies the index of equal distance iteration.

(ii) Use a set of K coefficients, $(\lambda^I)^{(n)}$ ($I = 0, K - 1$), to update the position vector of each replica I:

$$(\mathbf{r}^I)^{(n+1)} = (\mathbf{r}^I)^{(n)} + (\lambda^{I-1})^{(n)} \left(\frac{\partial \Delta I^{I-1}}{\partial \mathbf{r}^I} \right)^{(n)} + (\lambda^I)^{(n)} \left(\frac{\partial \Delta I^I}{\partial \mathbf{r}^I} \right)^{(n)} \quad (3)$$

These coefficients are going to be solved in step (iii). For a chain of $K + 1$ replicas, there are K segments along a chain, and K coefficients ($I = 0, K - 1$) are involved. Since we fix the two ends, eq 3 is applied from replicas I to replica $K - I$.

(iii) Solve $(\lambda^I)^{(n)}$ by setting the first order Taylor expansion of each of $((\Delta I^I)^{(n+1)} - \overline{\Delta l})$ ($I = 0, K - 1$) with respect to $(\lambda^I)^{(n)}$ to zero, i.e.,

$$-((\Delta I^I)^{(n)} - \overline{\Delta l}) = \sum_{j=I-1}^{I+1} \left(\frac{\partial \Delta I^I}{\partial \lambda^j} \right)_{\lambda^j s=0} (\lambda^j)^{(n)} \quad (4)$$

There are a total of K equations for K unknowns. Since each replica is only coupled to its nearest neighbors through

distance constraints, eq 4 is a set of tridiagonal algebraic equations, which can be solved using well-established methods.²⁹

(iv) If any of the values of $l((\Delta l^I)^{(n+1)} - \overline{\Delta l})$ ($I = 0, K - 1$) calculated from $\{(\mathbf{r}^0)^{(n+1)} \cdots (\mathbf{r}^K)^{(n+1)}\}$ (via eq. 3) is greater than tolerance, then repeat (ii–iii). A tolerance of 10^{-8} Å is used throughout this work.

During the iterations for solving eq 2, $\overline{\Delta l}$ remains fixed and the total length of a chain is left the same by steps i–iv; $\overline{\Delta l}$, though, can change freely during path optimization. Gradients of distances are employed in eq 3 as the basis set for adjusting the position of each replica, and λ^I designates the displacements along these vectors. Equal distance constraints are usually satisfied via steps ii–iv within 20–25 iterations, even during the initial stages of path optimization where large steps are often involved. The proposed scheme works equally well if the distances between replicas are defined via rms best-fit and demonstrate similar stability and efficiency for all molecular systems that we have tested so far. Generality to rms best-fit distance is a major advantage of using holonomic constraints instead of reparametrization schemes used in the string method. Compared to the use of restraint potentials, as in NEB, use of holonomic constraints guarantees high accuracy in the satisfaction of constraint equations via Lagrange’s multipliers, separating it from the optimization of the objective function. Furthermore, this approach can be generalized if variable density of replicas along a path is required by a given mechanism of distributing distances, for example, using more replicas around higher energy regions.

C. Objective Functions for Path Optimization. The proposed procedure for solving distance constraints can be applied with any objective function for path optimization. Several commonly employed functions for path optimization are briefly summarized, followed by the introduction of kinetic energy potentials and minimum Hamiltonian paths.

1. *Minimum Energy Path (MEP).* A solution of minimizing the total potential energy, $\sum_{I=0}^K V^I$, of a chain subject to equal distance constraints (eq 2) determines a MEP⁴ that satisfies:

$$(-\nabla_I V^I)(1 - \boldsymbol{\tau}^I \boldsymbol{\tau}^I) = 0 \quad (5)$$

$V^I = V(\mathbf{r}^I)$ is the potential energy of replica I , and $\boldsymbol{\tau}^I$ is the unit tangent vector of replica I . Under the framework of constrained optimization,²⁹ $\boldsymbol{\tau}^I$ are defined by the gradients of the constraint equations (eq 2) that involve replica I :

$$\boldsymbol{\tau}^I = \frac{\frac{\partial \Delta l^{I-1}}{\partial \mathbf{r}^I} + \frac{\partial \Delta l^I}{\partial \mathbf{r}^I}}{\left| \frac{\partial \Delta l^{I-1}}{\partial \mathbf{r}^I} + \frac{\partial \Delta l^I}{\partial \mathbf{r}^I} \right|} = \frac{\sum_{i=1}^N w_i (\mathbf{r}_i^{I+1} - \mathbf{r}_i^{I-1})}{\left| \sum_{i=1}^N w_i (\mathbf{r}_i^{I+1} - \mathbf{r}_i^{I-1}) \right|} \quad (6)$$

2. *Minimum Free Energy Path (MFEP).* Following Maragliano et al.,¹⁸ a MFEP is defined as a path on which the mean force at a finite temperature is parallel to the path, i.e.,

$$\langle -\nabla_I F^I \rangle_I (1 - \boldsymbol{\tau}^I \boldsymbol{\tau}^I) = 0 \quad (7)$$

F^I is the free energy for constraining the system at the configuration of replica I . Methods for optimizing MEPs can, thus, be generalized at a finite temperature to identify MFEPs,¹⁸ and the approach of using holonomic constraints can be readily applied. Details of identifying the MFEP with equal distance constraints and computing free energy profiles along a path will be discussed in future work.

3. *Dynamic Path.* Dynamic paths can be found by applying the least action principle³⁰ where the kinetic energy associated with each segment of a path can be defined as:

$$T^I = \sum_{i=1}^N \frac{1}{2} m_i \left(\frac{r_i^{I+1} - r_i^I}{\Delta t^I} \right)^2 = \frac{1}{2} \frac{M}{(\Delta t^I)^2} (\Delta l_m^I)^2 = \frac{1}{2} M (v_m^I)^2 \quad (8)$$

In eq 8, M is the total mass of the molecular system and Δl_m^I is the mass-weighted distance between replica I and $I + 1$ (see eq 1). Δt^I represents the time for the system to travel a distance of Δl_m^I , and $\Delta l_m^I / \Delta t^I$ is the mass-averaged root-mean-squared velocity, v_m^I . The direction of velocity vector is specified by the difference between two position vectors. Given a value of Δt^I , T^I gives the kinetic energy for connecting two neighboring replicas separated by Δl_m^I . Equation 8 is of the form of a potential energy function that is quadratic in Δl_m^I with $M/(\Delta t^I)^2$ as the force constant; therefore, T^I is referred to as kinetic energy potential. Assigning a force constant to a kinetic energy potential is equivalent to assigning a value of the time duration for connecting two replicas. If all of the atoms in a molecular system are included in defining the distance between replicas (eq 1), then T^I corresponds to the total kinetic energy of the system excluding translational and rotational degrees of freedom. For cases where a subset of the system is used to define the path, T^I corresponds to the kinetic energy of the subset.

When the distances between replicas are constrained to be equal, it is natural to use the total contour length, L , $L = \sum_{I=0}^{K-1} \Delta l_m^I$, instead of time as the independent variable for classical action.^{25,30} A discrete representation of action S_L for a chain with $K + 1$ replicas is^{25,30}

$$S_L = \int_0^L d\ell \sqrt{2(E - V)} \approx \Delta l_m \sum_{I=0}^K v_m^I = \frac{L}{K} \sum_{I=0}^K v_m^I \quad (9)$$

A chain of replicas that corresponds to a stationary point of S_L satisfies $(\delta S_L) / (\delta \mathbf{r}^I) = 0$, which leads to Newton’s equations of motion:³⁰

$$\left[\frac{m_i}{(\Delta t^I)^2} (r_i^{I+1} - 2r_i^I + r_i^{I-1}) \right] = [-\nabla_I V^I (1 - \boldsymbol{\tau}_m^I \boldsymbol{\tau}_m^I)]_i \quad (10)$$

In eq 10, $\boldsymbol{\tau}_m^I$ is the tangent vector associated with replica I ; the subscript m indicates that atomic masses are the weighting factors for defining distances between replicas (see eq 6). We also implicitly assume that all of the atoms in the molecular system are involved in defining the distances between replicas. The values of v_m^I in eq 9 may be determined by the work–energy theorem, and Δt^I in eq 10 can then be calculated as $\Delta l_m^I / v_m^I$. Finding stationary solutions of eq 9,

however, is very difficult,^{12,25} since they may include the minimum, maximum, or saddle points of S_L . Directly minimizing $|(\delta S_L)/(\delta \mathbf{r}^l)|$ requires the calculations of the Hessian matrices of replicas and is limited to relatively small molecular systems.¹²

4. *Minimum Hamiltonian Path (MHP)*. The sum of the kinetic energy potential (eq 8) and potential energy of a replica defines its Hamiltonian,³⁰ and the total Hamiltonian of a chain is

$$H^{\text{tot}} = \sum_{l=0}^{K-1} (T^l + V^l) = \sum_{l=0}^{K-1} H^l \quad (11)$$

Kinetic-energy potentials characterize the magnitude of inertia for keeping a path straight, and thus, providing a physically based approach for reducing curvatures along a path. If path optimization is conducted only with T^l , then a straight line connecting two configurations would result. The kinetic energy potential T^l in eq 8 is quadratic in Δl_m^l with $M/(\Delta t^l)^2$ as the force constant. Minimizing the total Hamiltonian of a chain subject to equal distance constraints gives a minimum Hamiltonian path (MHP) that satisfies:

$$\left[-\frac{m_i}{(\Delta t^l)^2} (r_i^{l+1} - 2r_i^l + r_i^{l-1}) \right] = [-\nabla_l V^l (1 - \boldsymbol{\tau}_m^l \boldsymbol{\tau}_m^l)]_i \quad (12)$$

The left-hand side of eq 12 is the same as that of eq 10, but with an opposite sign. Therefore, a MHP does not satisfy Newton's equation of motion, but such dynamics could be achieved if the system is coupled to an external bath. With all Δl_m^l 's kept at the same value, assigning a common value to all force constants gives constant kinetic energy potentials along a path, i.e., an isokinetic path.³¹ In this case, a MHP is also a minimum energy isokinetic path. Analogous to a MEP being the most probable path for an overdamped system,³² a minimum energy isokinetic path is most highly likely among paths that satisfy isokinetic conditions.³¹

The temperature associated with a kinetic energy potential can be calculated to characterize its magnitude.³¹ By scaling the force constant of kinetic energy potentials during path optimization, one may fix kinetic energy potentials at a desired value (or temperature) during path optimization. If kinetic energy potentials are set to zero, then the objective function becomes the total potential energy; a minimum energy path (MEP) is thus a 0 K MHP. Kinetic energy potentials also provide an estimation of the time scale of progression, $t^l = \Delta l_m / v_m^l$, associated with each segment. A low kinetic energy potential corresponds to a longer progression time. For a MEP, $v_m^l \rightarrow 0$ and $t^l \rightarrow \infty$.

Using kinetic energy potentials is, thus, a physically based approach to control curvature during path optimization. Highly kinked paths can result from having too many replicas⁶ or are due to the nature of objective functions. For example, when optimizing MFEP using the finite temperature string method, finite time molecular dynamics or Langevin dynamics are performed to calculate mean forces.¹⁶ As a result, statistical noise is inevitably involved and can cause instability and kinks during path optimization.¹⁶ In both cases, kinetic energy potentials can be used to maintain a

smooth path and stability during optimization. As shown later, the magnitude of kinetic energy potentials can be quantitatively controlled to minimize the effects on objective function.

D. Quantifying Accuracy of Reaction Path Optimization. On a MEP or MHP with isokinetics, the potential energy difference between replicas can be computed by applying the work–energy theorem, for which a discrete representation based on a second-order symmetric finite difference is

$$(\Delta V^l)^{\text{WET}} = \frac{1}{4} \sum_{J=l}^{l+1} (\nabla_J V^J) (\boldsymbol{\tau}_m^J) (\Delta l_m^{J+1, J-1}) \quad (13)$$

In eq 13, the length, over which $(-\nabla_l V^l) (\boldsymbol{\tau}_m^l)$ is exerted, is approximated by $\Delta l_m^{l+1, l-1}/2$. Equation 13 or, equivalently, the work–energy theorem, can be used to quantify the accuracy of reaction path optimization. For example, accuracy can be measured by the rms differences between $(\Delta V^l)^{\text{WET}}$ calculated via eq 13 and the calculated values of energy difference, $\Delta V^l = (V^{l+1} - V^l)$, $\delta(\Delta V)_{\text{rms}}$, as

$$\delta(\Delta V)_{\text{rms}} = \sqrt{\frac{\sum_{l=0}^{K-1} ((\Delta V^l)^{\text{WET}} - \Delta V^l)^2}{K}} \quad (14)$$

If a reaction path is accurate, then the resulting path will give a small value of $\delta(\Delta V)_{\text{rms}}$. If the algorithm of reaction path optimization is correct, then the value of $\delta(\Delta V)_{\text{rms}}$ is expected to decrease with the increase of the number of replicas used in a chain. As suggested in eq 13, factors that affect the value of $\delta(\Delta V)_{\text{rms}}$ include the degrees of freedom used in defining distances between replicas, the estimation of tangent vectors, and the calculation of distances between replicas. Since reaction path methods may differ in these properties, eqs 13 and 14 provide an unbiased and quantitative measure of accuracy.

Results and Discussion

To illustrate the enhancement in stability, efficiency, and robustness by employing equal distance holonomic constraints and kinetic energy potentials, they are applied to the path optimization of three molecular transitions: the $C_{7\text{eq}}$ -to- C_{ax} transition of an alanine dipeptide, the 4C_1 -to- 1C_4 transition of an α -D-glucopyranose, and the helix-to-sheet transition of a GNNQQNY heptapeptide. Since all-atom force fields are used, at least 20 replicas are required to represent transitions in all three cases for accurate reaction optimization (as measured by eqs 13 and 14). As such, both methods are found to be essential for the convergence of reaction path optimization. All calculations are performed with the CHARMM program,³³ based on which (version c35a2) we implemented the new developments. Simulation details will be described in each test case.

A. Isomerization of Alanine Dipeptide. The transition of alanine dipeptide ($\text{CH}_3\text{--CO--NH--CHCH}_3\text{--CO--NH--CH}_3$) between $C_{7\text{eq}}$ ($\phi = 77.3, \psi = 64.3$) and C_{ax} ($\phi = 78.2, \psi = -55.8$) isoforms (see Figure 1) is a standard test case for the development of computational methods that aim

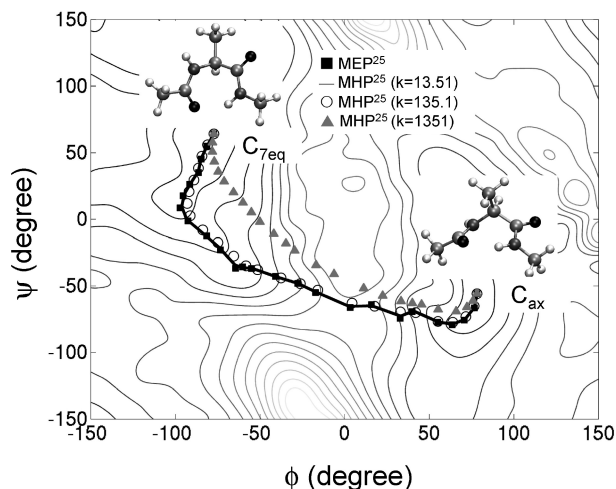


Figure 1. Pathways of the C_{7eq} -to- C_{ax} transition of an alanine dipeptide. Pathways are projected on a two-dimensional surface of backbone dihedral angles (ϕ : $C-N-C_{\alpha}-C$, ψ : $N-C_{\alpha}-C-N$). The contour plot is generated via minimized structures of alanine dipeptide restrained at different values of ϕ and ψ . Structures of the C_{7eq} and C_{ax} configurations are shown via a ball-and-stick representation. MEP stands for a minimum energy path, and MHP stands for a minimum Hamiltonian path; superscript indicates the number of replicas used in a chain. Values of the force constants of kinetic energy potentials (k 's) in kcal/mol/Å² are also specified. See text for discussion.

to find reaction paths. The CHARMM22 all-atom force field³⁴ with CMAP backbone dihedral angle corrections³⁵ is used for calculation without solvation, and all pair interactions between atoms are involved in calculating the potential energy. The initial configurations of replicas are generated by linearly interpolating in the ϕ and ψ space.

All atoms except methyl hydrogen are used to define the distances between replicas using eq 1 with a rms best-fit procedure to remove contributions from rigid body translation and rotation using atomic masses as weighting factors. Although the initial configurations of replicas generated via linear interpolation are far from being equally spaced based on mass-weighted rmsd, the proposed procedure for solving the constraint equations still manages to converge within 25 iterations for a chain of 25 replicas. The criterion of convergence is that the root-mean-squared difference in rms distance ($\delta(\Delta l)$) between replicas is less than 10^{-8} Å. Using steps i–iv as described earlier in section B, we also found that the required number of iterations for converging equal distance constraints is not sensitive to the number of replicas used in the chain. Compared to using restraint potentials, which corresponds to a single step of iteration, employing holonomic constraints requires additional cost in solving constraint equations to gain accuracy and to decouple maintaining equal distances from optimizing the objective function. Another widely used approach to maintain equal distances between replicas is reparametrization using interpolation functions such as cubic splines.^{15,17} If the rms best-fit procedure is not used to define the distance between replicas, then forming a continuous curve to connect replicas is straightforward and reparametrization is very efficient in equating the distances between replicas. Using the same

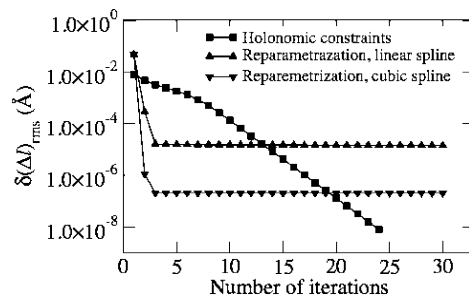


Figure 2. The rms difference in distances between replicas as a function of iteration for solving constraint equations and reparametrization. The initial path of the C_{7eq} -to- C_{ax} transition using 25 replicas is used as the testing case. See text for details.

initial path as described above, Figure shows that only four iterations of reparametrization are needed to reduce $\delta(\Delta l)$ to an accuracy determined by the interpolation functions. $\delta(\Delta l)$ using cubic splines increases the upper limit of accuracy by two orders of magnitude ($\delta(\Delta l) = 2 \times 10^{-6}$ Å) compared to that of using a linear interpolation ($\delta(\Delta l) = 2 \times 10^{-4}$ Å). On the other hand, since holonomic constraints only require the input of constraint equations, the accuracy of convergence does not depend on the choice of interpolation functions. Furthermore, if a rms best-fit procedure is applied to calculating distances between replicas, then using the holonomic constraints approach allows straightforward generalization, whereas the applicability of continuous curve approximation in the reparametrization approach requires further investigation, which is beyond the scope of this work. In both cases of using reparametrization shown in Figure 2, the rms best-fit procedure is not used in calculating distances between replicas. In conclusion, using holonomic constraints provides higher accuracy in satisfying the equal distance constraints and higher flexibility in defining the distance between replicas, as compared to other commonly used methods. Additional cost, though, is involved in solving the constraint equations. As shown in the following, decoupling the optimization of the objective function and maintaining equal distances between replicas using holonomic constraints play important roles in enhancing the stability and ensuring the accuracy of reaction path optimization. Furthermore, the percentage cost of solving constraint equations would decrease with the increase of system size and the complexity of energy calculations, for example, when ab initio methods are used.

The rms gradient of potential energy, $|\nabla V|_{rms}$, for a chain of 25 replicas during the course of ABNR (adopted-basis Newton–Raphson) minimization is shown in Figure 3 (left axis, log scale); also shown (right axis) is the number of iterations for solving constraint equations. Kinetic energy potentials are not used here. The superlinear convergence of ABNR minimization is clearly seen in Figure 3, with a convergence criterion of $|\nabla V|_{rms} < 10^{-5}$ kcal/mol/Å satisfied within 2 500 steps. By enforcing equal distance constraints, we illustrate that a MEP can be obtained without using restraint potentials, reparametrization, and/or force projections.

During the path optimization shown in Figure 3, the mass-weighted rmsd's ($rmsd_M$'s) between replicas are constrained

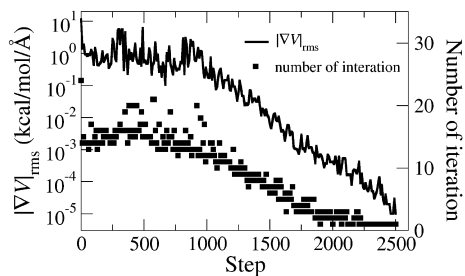


Figure 3. The rms gradient of potential energy $|\nabla V|_{\text{rms}}$ (labels on left, y-axis) and the number of iterations for solving the constraint equations of equal distances (labels on right, y-axis) during path optimization using ABNR. The chain contains 25 replicas of alanine dipeptide to connect $C_{7\text{eq}}$ and C_{ax} . Kinetic energy potentials are not used in path optimization.

to be equal. All atoms other than methyl hydrogen (13 out of 22) are used to define the distance between two replicas. As shown in Figure 3, the constraint equations can be satisfied within ~ 20 iterations via steps i–iv. As the optimization approaches convergence, the number of iterations also decreases since the magnitude of displacement per step also decreases. The value of rmsd_M of the converged MEP with 25 replicas is 0.103 \AA , and the total contour length of the MEP is, thus, 2.48 \AA ; the rmsd_M between $C_{7\text{eq}}$ and C_{ax} is 1.17 \AA . The 25 replica MEP (MEP^{25}) is projected onto a (ϕ, ψ) surface and shown in Figure 1. Without using holonomic constraints, reaction path optimization often fails to converge using restraint potentials alone. This tendency becomes more significant in the other two test cases since the underlying transition is more complicated.

Combining potential energy functions and kinetic energy potentials as the objective function of path optimization determines a minimum Hamiltonian path (MHP). MHPs of 25 replicas (MHP^{25}) with different magnitudes of kinetic energy potentials are computed to illustrate their effects on the resulting path; results are shown in Figure 1. With a $k \equiv M/(\Delta t)^2$ value of $13.51 \text{ kcal/mol/\AA}^2$, the resulting MHP (solid line in Figure 1) is indistinguishable compared to the MEP (solid squares). The rmsd_M between replicas on MHP^{25} ($k = 13.51 \text{ kcal/mol/\AA}^2$) is 0.1 \AA (the value for the MEP is 0.103 \AA), and the corresponding temperature of kinetic energy potentials has a very small value of 1.74 K . When k is increased to $135.1 \text{ kcal/mol/\AA}^2$, the resulting rmsd_M between replicas on the MHP becomes 0.092 \AA and the corresponding temperature is 14.8 K . The projected MHP^{25} ($k = 135.1 \text{ kcal/mol/\AA}^2$) on a (ϕ, ψ) surface (open circles in Figure 1) is also close to that of the MEP but is smoother near minima. When k is raised to $1351 \text{ kcal/mol/\AA}^2$, the rmsd_M between replicas becomes 0.068 \AA and the corresponding temperature is 80.61 K . From Figure 1, the deviation of MHP^{25} ($k = 1351 \text{ kcal/mol/\AA}^2$; solid triangles) from the MEP is clear: the path is straightened by kinetic energy potentials near local minima but to a much lesser extent near a saddle point.

The pathways shown in Figure 1 indicate that kinetic energy potentials affect more on regions with a flat potential energy surface, over which kinks tend to develop during path optimization.⁶ This trend can also be seen in the plots of potential energy profiles shown in Figure 4. Increasing the

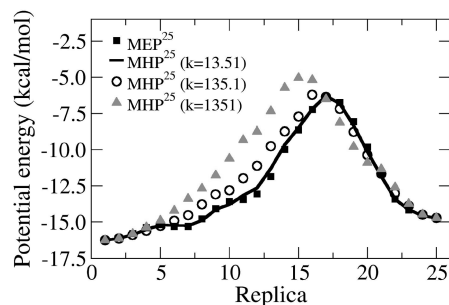


Figure 4. Profiles of potential energy on optimized pathways of the $C_{7\text{eq}}$ -to- C_{ax} transition of an alanine dipeptide in the gas phase. The x-axis specifies the index of the replica along a chain. MEP stands for a minimum energy path, and MHP stands for a minimum Hamiltonian path; superscript indicates the number of replicas used in a chain. Values of the force constants of kinetic energy potentials (k 's) in kcal/mol/\AA^2 are also specified.

strength of kinetic energy potentials shifts the potential energy profile to the left, since the $C_{7\text{eq}}$ basin (replica 1) is flatter than that of C_{ax} . The shoulder between replicas 5 and 8 on MEP^{25} also gradually disappears as k is increased from 13.51 to $1351 \text{ kcal/mol/\AA}^2$. The value of the highest potential energy along a path (replica 17 for MEP^{25} , MHP^{25} ($k = 13.51$ and $135.1 \text{ kcal/mol/\AA}^2$) and replica 15 for MHP^{25} ($k = 1351 \text{ kcal/mol/\AA}^2$)) does increase due to the presence of kinetic energy potentials but only becomes appreciable when $k = 1351 \text{ kcal/mol/\AA}^2$. ΔE^{max} , the difference in energy between the highest energy replica on MHP^{25} and MEP^{25} is 1.31 kcal/mol for $k = 1351 \text{ kcal/mol/\AA}^2$ but is only 0.07 and 0.001 kcal/mol for $k = 135.1$ and $13.51 \text{ kcal/mol/\AA}^2$, respectively. Differences in geometries due to kinetic energy potentials also follow a similar trend. $\text{rmsd}_M^{\text{max}}$, the rmsd_M between the highest energy replica on MHP^{25} and MEP^{25} , is 0.2 for $k = 1351$, 0.04 for $k = 135.1$, and 0.009 \AA for $k = 13.51 \text{ kcal/mol/\AA}^2$.

In summary, using 25 replicas to represent the $C_{7\text{eq}}$ -to- C_{ax} transition of an alanine dipeptide, kinetic energy potentials with $k < 150 \text{ kcal/mol/\AA}^2$ do not cause an appreciable difference between the resulting MHP and MEP ($\Delta E^{\text{max}} < 0.1 \text{ kcal/mol}$; $\text{rmsd}_M^{\text{max}} < 0.1 \text{ \AA}$). On the other hand, the number of steps for converging reaction path optimization ($|\nabla V|_{\text{rms}} < 10^{-5} \text{ kcal/mol/\AA}$) is much less (by 2–3 times) in the presence of kinetic energy potentials due to the reduced occurrence of kinks; results are shown in Figure 5. Even in the presence of a large number of replicas (up to 97), kinetic energy potentials that correspond to a temperature of ~ 2 – 5 K suffice to ensure stable path optimization. This level of kinetic energy potential ($< 5 \text{ K}$) causes a negligible difference in energetics ($\Delta E^{\text{max}} < 0.05 \text{ kcal/mol}$) and geometries ($\text{rmsd}_M^{\text{max}} < 0.05 \text{ \AA}$) compared to a MEP. Without kinetic energy potentials, path optimization becomes difficult to converge when a large number of replicas are used in a chain.

Therefore, kinetic energy potentials can be used to significantly enhance the stability of reaction path optimization and increase the efficiency by reducing the steps of optimization. These advantages can be exploited without causing appreciable perturbation in structures and energetics. All calculations presented above are obtained by setting a

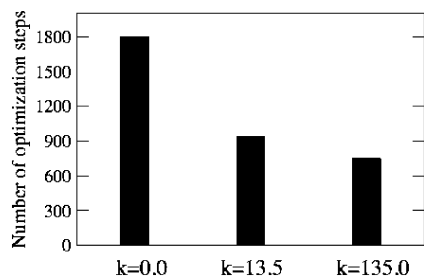


Figure 5. The required number of minimization steps for converging reaction path optimization ($\nabla V_{\text{rms}} < 10^{-5}$ kcal/mol/Å) using different values of force constants for kinetic energy potentials. The $C_{7\text{eq}}$ -to- C_{ax} transition of an alanine dipeptide using 25 replicas is used as the example.

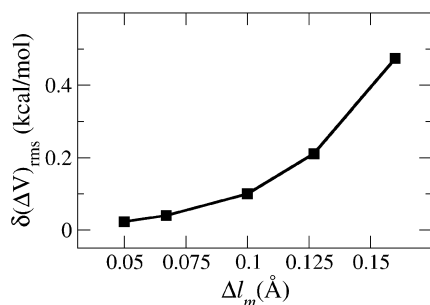


Figure 6. Rms error of using the work-energy theorem in computing potential energy differences between replicas via eq 13. Rms errors are plotted as a function of the rmsd_M between two neighboring replicas.

constant value of the force constants for kinetic energy potentials ($k \equiv M/(\Delta t)^2$, eq 8); an alternative approach is to fix the value of kinetic energy, $1/2M(\Delta l/\Delta t)^2$, or temperature by rescaling force constants during path optimization. When the same value of the force constant is used for all replicas and all distances between replicas are constrained to be equal, an isokinetic path results.³¹ In this case, a MHP corresponds to a minimum energy isokinetic path, which is also the most probable path among pathways that satisfy isokinetic conditions.³¹ Different schemes may also be designed to specify the force constants of kinetic energy potentials but will not be discussed in this work. Moreover, kinetic energy potentials can also be employed to systematically examine the effects of inertia on reaction pathways, as shown in Figure 1 and Figure 4.

A key parameter for chain-of-states methods is the number of states or replicas, and an important question is how many is considered enough. While it is difficult to predict a priori the required number of replicas for a transition, a useful metric is whether the work-energy theorem is satisfied by eqs 13 and 14. For reaction paths of the $C_{7\text{eq}}$ -to- C_{ax} transition of an alanine dipeptide with different numbers of replicas, $\delta(\Delta V)_{\text{rms}}$ is plotted as a function of Δl_m , the mass-weighted rms best-fit distance between replicas, in Figure 6. It can be seen that for $\Delta l_m = 0.1$ Å, $\delta(\Delta V)_{\text{rms}} = 0.1$ kcal/mol. Since a second-order scheme is employed in eq 13 to estimate $(\Delta V)_{\text{wet}}$, $\delta(\Delta V)_{\text{rms}}$ decreases quadratically with Δl_m as shown in Figure 6. The small value of $\delta(\Delta V)_{\text{rms}}$ for the case of alanine dipeptide also indicates that ignoring the hydrogen atoms of methyl groups in defining the distance between

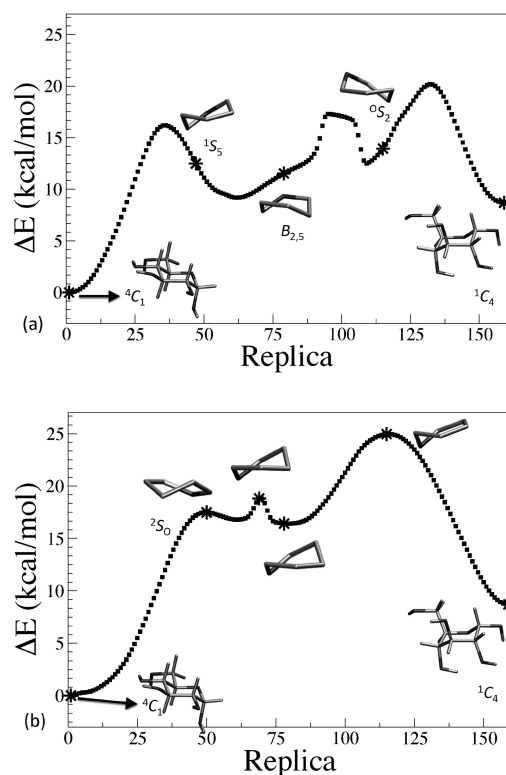


Figure 7. Potential energy profiles of two optimized pathways, (a) and (b), of the 4C_1 -to- 1C_4 transition of α -D-glucopyranose in the gas phase. The x-axis specifies the index of the replica. MHPs are optimized by using kinetic energy potentials with a low temperature (<5 K) for both paths. Structures of selected replicas are shown to illustrate the nature of the transition. See text for discussion.

replicas results in negligible error in computing the accumulated work along a reaction path.

To our knowledge, quantitative analysis of the accuracy of a reaction path via physically based approaches such as the work-energy theorem has not yet been conducted. Based on this criterion, the convergence of $\delta(\Delta V)_{\text{rms}}$ with the reduction of Δl_m in Figure 6 justifies the validity of our scheme of reaction path optimization, and we found that both the use of rms best-fit procedure and holonomic constraints are critical in ensuring the accuracy of path optimization. As described earlier, another factor that affects $\delta(\Delta V)_{\text{rms}}$ is the degrees of freedom used in defining distances between replicas and eqs 13 and 14 can be used to systematically examine whether an atom can be ignored in defining the reaction path or not. Figure 6 indicates that ignoring methyl hydrogen results in a negligible effect on the energy profile; ignoring any of the heavy atoms, though, results in a noticeable change (>1.0 kcal/mol) in energy profile.

B. 4C_1 -to- 1C_4 Transition of α -D-Glucopyranose. D-Glucose is an important monosaccharide in medicine, energy storage, and biology, and α -D-glucopyranose is an abundant anomer within this class of biomolecules. The properties and functions of polysaccharides are tightly related to the conformation of glucose monomers, and it is crucial to understand the relative energetics and interconversion between different conformational states.³⁶⁻³⁹ As shown in Figure 7, the most stable conformation of α -D-glucopyranose is the 4C_1 chair, and it can undergo structural transition to a

less stable 1C_4 chair. The 4C_1 -to- 1C_4 transition involves collective rearrangements of atoms in the pyranose ring and side-chain hydroxyl and the methylhydroxyl groups and, thus, requires a robust reaction path method to analyze mechanisms. Holonomic equal distance constraints and kinetic energy potentials are applied to simulate the pathways of 4C_1 -to- 1C_4 transition of an α -D-glucopyranose molecule in the gas phase. Since a large magnitude of forces are involved in the transition, a large number of replicas is required. Both methods of employing holonomic constraints and kinetic energy potentials are found to be necessary for reaction path optimization to converge.

The CHARMM35 carbohydrate force field⁴⁰ is used to describe the potential energy surface of α -D-glucopyranose conformation without using a cutoff for nonbonded interactions. The minimized *tg* isomers^{39,41} of 4C_1 and 1C_4 are used as the initial and final replicas. The difference in energy between 4C_1 and 1C_4 is 8.7 kcal/mol, close to the result of 7.93 kcal/mol obtained from DFT calculations at the B3LYP/6-311++G** level.⁴² All atoms are used to define the distance between replicas via mass-weighted rmsd best-fit. Since a significant number of hydrogen atoms are involved, a large number of replicas are needed to resolve the structural arrangements during transition. From an initial path generated via linear interpolation, a path of 159 replicas is optimized and shown in Figure 7(a). The rmsd_M between replicas is 0.02 Å, and the magnitude of kinetic energy potentials used for path optimization is 3 K. A small rmsd_M of 0.02 Å between replicas ensures a small $\delta(\Delta V)_{\text{rms}}$ of 0.04 kcal/mol.

Based on the conformational nomenclature of six-membered rings,⁴¹ the path in Figure 7(a) follows a route of 4C_1 - 1S_5 - $B_{2,5}$ - 0S_2 - 1C_4 (*S*: skew; *B*: boat), and the structures of selected conformations are highlighted in Figure 7(a). Although rotations of side-chain atoms are also involved, the first and third barriers in Figure 7(a) mainly correspond to rearrangements of the six-member ring. The highest energy state is 20.2 kcal/mol higher in energy compared to 4C_1 , consistent with the results of DFT calculations.³⁷ The second barrier of Figure 7(a), on the other hand, is associated with the rearrangements of hydroxyl and methylhydroxyl groups without noticeable changes in the ring conformation. This result indicates that side-chain atoms are explicitly involved in the activation process of the 4C_1 -to- 1C_4 transition.

Since a $B_{2,5}$ -like intermediate is found in the path shown in Figure 7a, we examine whether a path with ${}^{2,5}B$ boat intermediate can also be identified. Using ${}^{2,5}B$ boat as an intermediate for linear interpolation, an initial path is generated to test this mechanism. The optimized path with 159 replicas is shown in Figure 7b, with an rmsd_M of 0.02 Å between replicas. The magnitude of kinetic energy potentials used for path optimization is 3 K. Although replicas with a 2S_0 -like structure are found along the path shown in Figure 7b (in contrast to the 0S_2 structure in the path of Figure 7a), structures similar to ${}^{2,5}B$ boat are not identified due to steric effects of the methylhydroxyl group. This result further affirms the importance of side-chain atoms in conformational transitions of glucopyranose. The path of Figure 7b has a higher barrier of 25.0 kcal/mol compared to that of the path in Figure 7a.

Combing holonomic equal distance constraints and kinetic energy potentials allows the identification of smooth pathways that use all 24 atoms of α -D-glucopyranose to define distances between replicas. Three barrier-crossing events are involved in both pathways shown in Figure 7. Activation processes include rearrangements of side-chain atoms in addition to those of the pyranose ring. These results highlight the complexity of glucopyranose conformational changes and the requirement of a stable and efficient reaction path optimization scheme to analyze mechanisms. As such, this test case has not been widely attempted by other methods of reaction path optimization. Other pathways of the 4C_1 -to- 1C_4 transition can be found by generating alternative initial paths via different interpolation strategies (currently ongoing work). Pathways obtained from molecular mechanical force fields may then be refined via ab initio methods.

C. Helix-to-Sheet Transition of a GNNQQNY Heptapeptide. The GNNQQNY heptapeptide is in an N-terminal prion-determining domain of Sup35, a prion-like protein in yeast that forms fibrillar amyloid assemblies.⁴³ Prions (proteinaceous infectious particles) are the cause of several neurodegenerative diseases and are formed by a prion protein whose conformation undergoes transition from the normal form to a misfolded form.⁴⁴⁻⁴⁶ The GNNQQNY heptapeptide exhibits the amyloid properties of full-length Sup35,⁴³ and the helix-to-sheet transformation of this has crucial implications in the fibril formation of the full-length prion protein.⁴⁴⁻⁴⁶

The helix form of the GNNQQNY peptide is constructed according to the canonical α -helical geometry, and the β -sheet form is obtained from X-ray crystallography (PDB code 1YJP).⁴³ The CHARMM22 all-atom protein force field with CMAP corrections was used with the FACTS implicit solvent model⁴⁷ to describe the potential energy surface of the GNNQQNY heptapeptide. Both α -helix and β -sheet structures are stable local minima, with the α -helix structure being 9.5 kcal/mol lower in potential energy. Including all atoms except the N-terminal hydrogen atoms to estimate the structural difference, the mass-averaged rms difference (rmsd_M) between the α -helix and β -sheet structures of the GNNQQNY peptide is 5.42 Å. It is expected that a pathway of the helix-to-sheet transition would involve multiple intermediate states and saddle points, thus, providing a stringent test for the stability and efficiency of reaction path optimization.

An initial path (~ 20 replicas) is first generated by restraining the molecule to different values of $((\text{rmsd}_M\text{-to-helix})-(\text{rmsd}_M\text{-to-sheet}))$ to connect the two structures. Even though the initial configurations are far from equally spaced, steps i-iv still manage to converge, indicating the robustness of the proposed scheme. As in the case of alanine dipeptide, an rmsd_M of 0.1 Å is needed to reduce the value of $\delta(\Delta V)_{\text{rms}}$ (eq 14) to smaller than 0.1 kcal/mol and is used as the criterion for determining the number of replicas in a chain. Path optimization starts with higher kinetic energy potentials (path temperature ≈ 15 K), and their magnitude is then reduced to a temperature lower than 3 K, which ensures stable path optimization with minimal effects on the potential energy profile as compared to a MEP. If the rmsd_M of a converged path is larger than 0.1 Å, a new replica is inserted

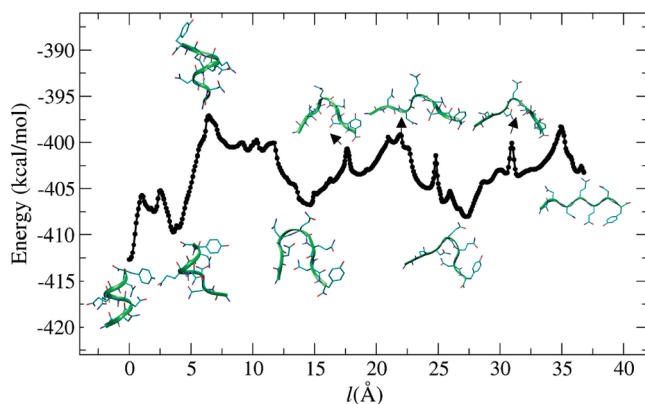


Figure 8. Potential energy profiles of an optimized path of the helix-to-sheet transition of a solvated GNNQQNY heptapeptide as a function of contour length. The path contains a total of 464 replicas. Structures of selected replicas are shown to illustrate the nature of the transition. See text for discussion and the strategy for optimization.

in between each pair of replicas to double the number of segments until the rmsd_M between replicas of a converged path is around 0.1 Å.

Although steps i–iv can equally space a chain of hundreds of replicas (for the GNNQQNY peptide, the maximum number of replicas tested is 1 024), the number of steps required for path optimization also increases. Therefore, we use up to 100 replicas in a chain for path optimization. If doubling the number of segments along a chain results in more than 100 replicas, then the chain is divided into two segments at the location of a stable intermediate state. For the helix-to-sheet transition, this divide-and-conquer strategy results in a total of seven segments and a total of 456 replicas of the jointed chain. Potential energies of all replicas along the jointed chain and structures of selected intermediates and saddle points are shown in Figure 8.

For the path shown in Figure 8, the α -helix starts to unwind from the N-terminal and then from the C-terminal end, along which the replica of the highest energy (15.5 kcal/mol higher than the α -helix) is located. The peptide then goes through a series of configurations with coil and hairpin-like structures to reach to the β -sheet structure. Multiple intermediates and saddle points are identified; the total contour length (measured by rmsd_M) for the chain is 36.7 Å, ~ 7 times larger than the rmsd_M between the α -helix and β -sheet structures (5.42 Å).

The computed barrier of 15.5 kcal/mol for the helix-to-sheet transition of the GNNQQNY peptide is lower than the activation energy (20–40 kcal/mol) of the helix-to-sheet transition of a longer segment (142-residue) of a mouse prion protein estimated by CD spectra at different temperatures.⁴⁶ By combining holonomic equal distance constraints and kinetic energy potentials with a divide-and-conquer strategy, we identify a pathway whose energetics are consistent with experimental measurements. This result has been difficult to achieve by applying chain-of-states methods^{46,48} due to the limited stability and efficiency discussed earlier and demonstrates that the methods developed in this work can be used to enable the convergence of reaction path optimization of complex molecular systems.

For complicated transitions such as the helix-to-sheet transition, there may be multiple pathways to connect two states. While sampling reaction pathways is not the focus of this work, it is important to emphasize that holonomic equal distance constraints and kinetic energy potentials combined with the divide-and-conquer approach provide an efficient framework for exploring different initial paths and mechanisms. Initial paths can be generated using different strategies, for example, from fast pulling simulations along different order parameters or using different combinations of internal coordinates for interpolation. With the enhanced robustness and efficiency using the two new methods presented in this work, high-quality initial paths are not required for reaction path optimization. Alternative pathways can also be found by initiating molecular dynamics simulations from an optimized path, following the strategy of transition-path sampling.^{49–51} Applications of these methods of helix-to-sheet transition of the GNNQQNY peptide and other conformational changes of biomolecules are ongoing.

Conclusion

Two new methods are developed in this work to enable and accelerate the convergence of reaction path optimization using a chain of replicas, especially for transitions occurring in complex molecular systems. First, the requirements of maintaining equal distances between replicas are achieved by holonomic constraints to transform the finding of a reaction path to a problem of constrained optimization. This approach allows precise implementation of analytical theory without imposing ad hoc procedures such as adding restraint potentials or reparametrization. For finding a minimum energy or minimum free energy path, force projections are not required and fast-converging optimization schemes such as quasi-Newton methods can be readily applied. A simple, fast, and robust scheme is developed for solving the constraint equations that demonstrates high stability and efficiency even if complex molecular systems and many replicas are involved. This scheme also supports the use of a rms best-fit procedure to measure the distance between replicas.

Second, we propose to use kinetic energy potentials to prevent the development of kinks during path optimization and introduce a new objective function, the total Hamiltonian of a chain of replicas, for reaction path optimization. Our results indicate that by reducing the development of kinks, the number of steps for optimizing a path can be reduced significantly, by 2–3 times. By using low-temperature kinetic energy potentials, this enhancement in stability and efficiency can be acquired without causing appreciable differences in structures and energetics between minimum energy and minimum Hamiltonian paths.

To quantify the accuracy of reaction path optimization, a method is developed based on the work–energy theorem (eqs 13 and 14). The results of this analysis (Figure 6) justify the validity of our scheme of reaction path optimization. To our knowledge, a quantitative assessment of the accuracy of reaction path optimization has not yet been conducted. The analysis proposed in this work has solid theoretical foundation and can be used to test different strategies and

hypothesis used in reaction path optimization, such as the degrees of freedom, that should be used for defining distances between replicas.

To illustrate the enhanced stability, efficiency, and robustness of reaction path optimization by applying the methods developed in this work, they are applied to three test cases that involve rugged potential energy surfaces and high dimensionality: the C_{7eq} -to- C_{ax} transition of alanine dipeptide, the 4C_1 -to- 1C_4 transition of an α -D-glucopyranose, and the helix-to-sheet transition of a GNNQQNY heptapeptide. All-atom force fields are used in all three cases. By employing holonomic constraints to maintain equal distances between replicas and kinetic energy potentials to prevent kinks, convergence of reaction path optimization can be achieved even when a large number of replicas (>100) are used. Furthermore, convergence of path optimization is superlinear and is achieved by using an adopted-basis Newton–Raphson (ABNR) method. In this case, the application of fast-converging optimization schemes is straightforward since force projections are not involved due the use of holonomic constraints. The enhanced efficiency of reaction path optimization allows the identification of pathways of complex molecular systems whose barriers of transition lie in physiologically relevant range and can be compared with experimental measurements.

Finally, we observe that the optimized pathways in all three test cases demonstrate rearrangements of distinct parts of a molecular system, indicating the difficulty of using a few simple order parameters to describe the underlying transition. Using a chain of replicas to define a one-dimensional curve appears to be a general framework to overcome this difficulty. The methods developed in this work are easy to implement in other chain-of-states methods and can be used to enable the applications of this strategy to a broader range of molecular systems.

Acknowledgment. This work is supported by the National Institutes of Health through a training grant (no. T32GM008352), National Science Foundation through the graduate research fellowship program, and the University of California, Berkeley.

References

- (1) Miller, W. H.; Handy, N. C.; Adams, J. E. *J. Chem. Phys.* **1980**, *72*, 99–112.
- (2) Elber, R.; Karplus, M. *Chem. Phys. Lett.* **1987**, *139*, 375–380.
- (3) Czerminski, R.; Elber, R. *Int. J. Quantum Chem.* **1990**, *38*, 167–186.
- (4) Olender, R.; Elber, R. *J. Mol. Struct.* **1997**, *398*, 63–71.
- (5) Jónsson, H.; Mills, G.; Jacobsen, K. W. In *Classical and Quantum Dynamics in Condensed Phase Simulations*; Berne, B. J., Ciccotti, G., Coker, D. F., Eds.; World Scientific: Singapore, 1998, p 185.
- (6) Henkelman, G.; Jónsson, H. *J. Chem. Phys.* **2000**, *113*, 9978–9985.
- (7) Henkelman, G.; Uberuaga, B.; Jónsson, H. *J. Chem. Phys.* **2000**, *113*, 9901–9904.
- (8) Chu, J.; Trout, B.; Brooks, B. *J. Chem. Phys.* **2003**, *119*, 12708–12717.
- (9) Trygubenko, S.; Wales, D. *J. Chem. Phys.* **2004**, *120*, 2082–2094.
- (10) Crehuet, R.; Field, M. J. *J. Chem. Phys.* **2003**, *118*, 9563–9571.
- (11) Huo, S. H.; Straub, J. E. *J. Chem. Phys.* **1997**, *107*, 5000–5006.
- (12) Passerone, D.; Parrinello, M. *Phys. Rev. Lett.* **2001**, *87*.
- (13) Burger, S. K.; Yang, W. T. *J. Chem. Phys.* **2006**, *124*, 054109.
- (14) Burger, S. K.; Yang, W. T. *J. Chem. Phys.* **2007**, *127*, 164107.
- (15) E, W.; Ren, W.; Vanden-Eijnden, E. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2002**, *66*, 052301.
- (16) E, W.; Ren, W.; Vanden-Eijnden, E. *J. Phys. Chem. B* **2005**, *109*, 6688–6693.
- (17) E, W.; Ren, W.; Vanden-Eijnden, E. *J. Chem. Phys.* **2007**, *126*, 164103.
- (18) Maragliano, L.; Fischer, A.; Vanden-Eijnden, E.; Ciccotti, G. *J. Chem. Phys.* **2006**, *125*, 024106.
- (19) Maragliano, L.; Vanden-Eijnden, E. *Chem. Phys. Lett.* **2007**, *446*, 182–190.
- (20) Khavrutskii, I. V.; Arora, K.; Brooks, C. L. *J. Chem. Phys.* **2006**, *125*, 7.
- (21) Woodcock, H.; Hodoscek, M.; Sherwood, P.; Lee, Y.; Schaefer, H.; Brooks, B. *Theor. Chem. Acc.* **2003**, *109*, 140–148.
- (22) Elber, R. *Curr. Opin. Struct. Biol.* **2005**, *15*, 151–156.
- (23) Karplus, M.; Kuriyan, J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6679–6685.
- (24) Sheppard, D.; Terrell, R.; Henkelman, G. *J. Chem. Phys.* **2008**, *128*, 10.
- (25) Elber, R.; Ghosh, A.; Cardenas, A. *Acc. Chem. Res.* **2002**, *35*, 396–403.
- (26) Elber, R.; Meller, J.; Olender, R. *J. Phys. Chem. B* **1999**, *103*, 899–911.
- (27) Kabsch, W. *Acta Crystallogr., Sect. A: Cryst. Phys., Diffraction, Theor. Gen. Crystallogr.* **1976**, *32*, 922–923.
- (28) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford: New York, 1987.
- (29) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in FORTRAN*; 2 ed.; Cambridge University Press: New York, 1992.
- (30) Landau, L. D.; Lifshitz, E. M. *Mechanics*; Elsevier: Oxford, U.K., 1976.
- (31) Evans, D. J.; Morriss, G. P. *Statistical Mechanics of NonEquilibrium Liquids*; Cambridge University Press: New York, 1990.
- (32) Vanden-Eijnden, E.; Heymann, M. *J. Chem. Phys.* **2008**, *128*, 061103.
- (33) Brooks, B. R.; Brucoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187–217.
- (34) 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.
- (35) Mackerell, A. D.; Feig, M.; Brooks, C. L. *J. Comput. Chem.* **2004**, *25*, 1400–15.
- (36) Biarnes, X.; Ardevol, A.; Planas, A.; Rovira, C.; Laio, A.; Parrinello, M. *J. Am. Chem. Soc.* **2007**, *129*, 10686–10693.
- (37) Momany, F. A.; Appell, M.; Willett, J. L.; Bosma, W. B. *Carbohydr. Res.* **2005**, *340*, 1638–1655.
- (38) Barrows, S. E.; Dulles, F. J.; Cramer, C. J.; French, A. D.; Truhlar, D. G. *Carbohydr. Res.* **1995**, *276*, 219–251.
- (39) Barrows, S. E.; Storer, J. W.; Cramer, C. J.; French, A. D.; Truhlar, D. G. *J. Comput. Chem.* **1998**, *19*, 1111–1129.
- (40) Guvench, O.; Greene, S. N.; Kamath, G.; Brady, J. W.; Venable, R. M.; Pastor, R. W.; Mackerell, A. D. *J. Comput. Chem.* **2008**, *29*, 2543–64.
- (41) *Eur J Biochem* IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN) **1980**, *111*, 295–8.
- (42) Appell, M.; Strati, G.; Willett, J. L.; Momany, F. A. *Carbohydr. Res.* **2004**, *339*, 537–51.
- (43) Balbirnie, M.; Grothe, R.; Eisenberg, D. S. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 2375–80.
- (44) Baskakov, I. V.; Legname, G.; Prusiner, S. B.; Cohen, F. E. *J. Biol. Chem.* **2001**, *276*, 19687–90.
- (45) Pan, K. M.; Baldwin, M.; Nguyen, J.; Gasset, M.; Serban, A.; Groth, D.; Mehlhorn, I.; Huang, Z.; Fletterick, R. J.; Cohen, F. E. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 10962–6.
- (46) Lipfert, J.; Franklin, J.; Wu, F.; Doniach, S. *J. Mol. Biol.* **2005**, *349*, 648–58.
- (47) Haberthuer, U.; Cafilisch, A. *J. Comput. Chem.* **2008**, *29*, 701–715.
- (48) Koslover, E. F.; Wales, D. J. *J. Chem. Phys.* **2007**, *127*.
- (49) Bolhuis, P.; Chandler, D.; Dellago, C.; Geissler, P. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291–318.
- (50) Dellago, C.; Bolhuis, P.; Geissler, P. *Adv. Chem. Phys.* **2002**, *123*, 1–78.
- (51) Dellago, C.; Bolhuis, P. G. *Top. Curr. Chem.* **2007**, *268*, 291–317.

CT9001398

Convergence and Heterogeneity in Peptide Folding with Replica Exchange Molecular Dynamics

Edmund Lin and M. Scott Shell*

*Department of Chemical Engineering, University of California Santa Barbara, 552
University Road, Santa Barbara, California 93106-5080*

Received March 13, 2009

Abstract: Replica exchange molecular dynamics (REMD) techniques have emerged as standard tools for simulating peptides and small proteins, in part, to evaluate the accuracy of modern classical force fields for polypeptides. However, it often remains a challenge to unambiguously discriminate force field flaws from simulations that do not reach equilibrium convergence. Here, we examine closely the convergence behavior of REMD runs for 14 test alpha and beta peptide systems, using an AMBER force field with a generalized Born/surface area implicit solvation model. Somewhat surprisingly, we find that convergence times can be quite large compared to the time scales reached in many earlier REMD efforts, with some short peptides requiring up to 60 ns of run time (per replica). Moreover, we detect a high degree of run-to-run heterogeneity, finding that REMD runs of the same peptide seeded with different initial velocities can exhibit a range of fast- and slow-folding behavior. By increasing the number of swap attempts per REMD cycle, we are able to reduce heterogeneity by diminishing the presence of slower-folding trajectories. Finally, we notice that convergence often can be signaled by a spike in the population of the most populated configurational cluster — a metric that is independent of the native structure. These results suggest that the systematic application of long runs, multiple trials, and convergence indicators may be important in future folding studies and in force field development efforts.

Introduction

The prediction of the three-dimensional native structures of proteins from their sequences remains a prominent challenge. It has been particularly challenging to make such predictions without bioinformatics scoring functions or templates, and instead using purely “physics-based” methods; these techniques involve canonical sampling and classical physiochemical energy functions and solvation models designed to reproduce true interatomic energetics. Such approaches have met two limitations: (1) computational challenges in sampling the vast conformational spaces and concomitant long folding time scales of proteins, and (2) potential deficiencies in current force field parametrizations. In order to delineate among these potential sources of error, it is critical to identify whether a simulation has reached equilibrium convergence, as even accurate force fields will not

correctly fold a protein if the simulation time scale is less than the intrinsic folding time. Therefore, the ability to assess and reach convergence in folding simulations is of major importance.

Two complementary approaches have emerged as standards for increasing simulation convergence and efficiency. First, implicit solvation models reduce the computational overhead of explicit water simulations and remove solvent viscosity that slows exploration of the conformational landscape. In particular, ongoing efforts have refined fast generalized Born/surface area (GBSA) techniques that generate approximate solutions to the continuum solvent electrostatics¹ with a nonpolar term proportional to solute surface area.² Second, the parallel replica exchange molecular dynamics (REMD) technique of Sugita and Okamoto speeds convergence relative to brute force molecular dynamics.³ In REMD, a number of copies (“replicas”) of a system are evolved simultaneously, each at a different temperature

* Corresponding author. E-mail: shell@engineering.ucsb.edu.

spanning that of interest (e.g., physiological temperature) to a heated state more conducive to traversing free energy barriers. Periodic swaps of neighboring replicas enable conformations to heat up and cool down and are performed in such a way as to maintain an overall Boltzmann-weighted ensemble at each temperature.

In this work, we assess the convergence properties of REMD for a number of peptides modeled with an atomic force field and GBSA solvation. We demonstrate that, even for very short systems, convergence times can be notably longer than those suggested by earlier studies. Moreover, we detect substantial run-to-run heterogeneity in simulations of the same helical peptide, with trajectory folding rates varying by up to a factor of 4–5. We then perform folding simulations of a variety of α -helical and β -hairpin peptides of 11–25 amino acids in length, finding convergence times as great as 60 ns (4.0 aggregate CPU-years). Our results suggest several strategies for assessing and accelerating convergence, and they emphasize the importance of well-converged simulations in the evaluation of force field performance.

Our work follows a long history of computational folding studies. Early efforts used Monte Carlo methods^{4–6} and single trajectory^{7–12} and multitrajectory molecular dynamics with massive distributed computing^{13–15} to examine small peptide systems with well-defined secondary structures and proteinlike folding behavior.¹⁶ Initial applications of the REMD technique employed explicit^{17–19} and implicit water²⁰ simulations that were limited to very short times (≤ 10 ns). Some of these studies suggested substantial force field sensitivity^{8,11,19,21} and raised questions about the balance of α/β energetics, particularly in the backbone torsions.^{6,13,17,22}

In recent years, however, REMD simulations using GBSA solvation have typically been performed for 10s or 100s of ns, and many of the earlier studies would now be considered much shorter than modern standards. Rao and Caffisch, for example, performed a 1 μ s simulation of a 20-residue β -sheet peptide and observed 312 folding and unfolding events.²³ In a series of separate force-field refinement studies, Jang et al. used 20–130 ns runs to fold a variety of α , β , and α/β peptides up to 28 residues in length.^{24–28} In particular, their work demonstrated the need for long runs in converging slow or frustrated folders;^{25–27} their refined force field²⁵ (a modification of AMBER 99²⁹) ultimately successfully folded de novo a number of systems to sub 3 Å accuracy.²⁷ Lei et al. folded the villin headpiece, a 35-residue helical miniprotein, to <2 Å accuracy using 200 ns REMD simulations and the AMBER 03 force field. The same group recently folded the 45-residue helical protein A to high accuracy, requiring a 200 ns effort.³⁰

The issues of efficiency and convergence in biomolecular simulations have been addressed in a number of ways. Several groups have argued that REMD speeds convergence relative to conventional molecular dynamics simulations for the equivalent amount of computational effort.^{23,30–33} Comparing conventional to replica exchange simulations, authors have found speedup factors of 14–72,³¹ 4–8,³² and ~ 5 ,³⁰ generally decreasing with the size of the protein. Daura and co-workers suggested that convergence could be signaled by

a leveling in the number of clusters taken from a sliding-window configurational clustering analysis,^{34,35} and Lyman and Zuckerman later extended this approach with a faster clustering algorithm and the additional consideration of cluster populations.³⁶ Recently, Abraham and Gready developed a simple model for assessing the “mixing efficiency” of REMD simulations related to the rate at which replicas are able to decorrelate from temperatures during exchanges.³⁷ This work suggests an absolute minimum simulation length in terms of system size and number of replicas, which is computed under the assumption that successive exchange attempts are completely decorrelated; while their estimated minimum times are actually small relative to typical efforts in the literature, correlations are likely to greatly increase such estimates.

Recent simulation efforts have attempted to clarify longstanding issues about potential flaws in modern force fields, in part, by achieving better simulation convergence than early studies. Studies have addressed the accuracy of backbone torsional energetics related to the appropriate balance between α -helical and β -sheet secondary structure motifs. Several groups have performed direct comparisons of simulations with infrared spectroscopy and NMR experiments.^{38–40} Importantly, these efforts identified an increased helical propensity in early force fields like AMBER 94^{38,39} and even suggested persistent problems with more recent parametrizations,³⁹ an issue that remains, perhaps in part, due to a greater focus on helical miniproteins in recent extensive REMD folding efforts. Still, several groups have pursued modification of the backbone torsional terms to improve accuracy and reduce helicity.^{24,25,38,41,42} Hu et al. also showed that a quantum mechanical treatment improved the prediction of Ramachandran distributions in capped monopeptides relative to several classical force fields and suggested problems with the latter’s parametrization by in vacuo rather than solvated ab initio calculations.⁴³ Some more recent parametrizations have attempted to address this problem with ab initio calculations designed to better mimic solvent polarization.⁴⁴

GBSA implicit solvation models have also been scrutinized as a potential source of error. Earlier studies by Zhou and co-workers found that free energy surfaces of a hairpin peptide were distorted by implicit solvation when compared to explicit.^{45,46} Several groups have suggested that a major contributor to this problem has been salt bridges, which are overstabilized by the GBSA model.^{46–49} In addition, Roe et al. have found that implicit solvation increases the helical propensity of force fields through its effect on backbone hydrogen-bonding strengths.⁴⁹ Two approaches have been pursued to correct these flaws. Okur et al. propose hybrid explicit/explicit REMD simulations that use a solvation shell of explicit water near a protein with an implicit approach beyond. Groups have also pursued modification of the intrinsic GB radii of various atom types.^{25,48} In both approaches, the authors found improved agreement with experimental structures and interaction strengths.

Still, it remains challenging to unambiguously decouple force field assessment and simulation convergence in REMD folding simulations. In particular, for force field studies that

do not yield good agreement with experiments, it is essential to rule out the possibility of unconverged results and too short run times. Three issues potentially affect this consideration: (1) **heterogeneity** – Molecular dynamics studies have shown highly heterogeneous folding behavior in even fast-folding peptides.⁵⁰ These results suggest REMD simulations can also be characterized by significant run-to-run variances. What is the magnitude of this heterogeneity, to what extent are multiple trials required to attain an accurate picture, and can such variances be reduced? (2) **fluctuations** – Short peptides typical of the lengths currently amenable to direct folding studies are likely to experience significant conformational fluctuations when compared to globular proteins.³⁶ Studies already suggest that converged simulations typically give average root-mean-square deviation (rmsd) values in the range of 2–4 Å.^{12,17,27,30,33,36} What is a reasonable baseline for comparing structures in these simulations when fluctuations are present? (3) **convergence times** – Some studies have found REMD folding times that significantly exceed the short, sub 10 ns run durations of early studies.^{25–27} What is the range of convergence times that one might expect for short peptides, and are there any metrics, independent of knowledge of the native structure, that can signal convergence?

It is the goal of this work to provide some perspectives on these questions using extensive REMD simulations. Here, we assess the behavior of REMD simulations performed with a modern and classical interatomic force field in folding a wide range of short peptides taken from the PDB. In particular, we focus on the magnitude, breadth, and heterogeneity of convergence time scales across the spectrum of sequences studied, performing long simulation runs where necessary to provide systems with ample sampling time. For a force field, we use an AMBER variant with an implicit solvation model that we and our co-workers previously found to correctly stabilize the folded structures of small peptides and proteins.^{51–53} We describe two studies, a detailed, multitrial examination of the short helical EK peptide and a broader investigation of convergence in a variety of α and β peptide systems.

Methods

We examined 11 peptides taken from the Protein Databank (PDB) and 3 suggested by earlier folding studies.^{19,21,26,27} For those taken from the PDB, we filtered for structures that were classified as peptide in the SCOP (structural classification of proteins) scheme,⁵⁴ had well-defined secondary structure elements, were solved under normal aqueous solution conditions, and had, at most, a single disulfide bond. Among the other three peptides, the helical EK peptide has been studied by Baldwin and co-workers⁵⁵ using circular dichroism and was found to have partial helicity ($\sim 40\%$) at 0 °C; we took its native structure to be an energy-minimized ideal helix for this sequence. The 15- β peptide was engineered by ref 56 and studied earlier by Kim et al.;²⁷ we obtained the unpublished structure from the authors in ref 56. The C-peptide, a helix from the C-terminal region of ribonuclease A originates from the work of Baldwin and co-workers,⁵⁷ and it has been extensively studied by

simulation;^{4,19,21,26,47} we take its native structure from the crystal structure of ribonuclease A.

We simulated each peptide using REMD. Our replica temperatures spanned 270–600 K and swaps were attempted every 20 ps of MD simulation. The number of replicas varied from 15–25 and were adjusted so that swap attempts were accepted on average with ~ 40 –50% probability. MD trajectories were produced using the sander program in the AMBER 9 software suite, and replica swapping and data collection were performed using a custom Python wrapper. Unless indicated otherwise in the PDB files or reference experimental data, all peptides were simulated without capping groups (i.e., with charged termini). Simulations were initiated with peptides in extended conformations, and trajectory frames were recorded every 1 ps. Any disulfide bonds present in the native structure were applied to the peptides at the start of the simulations.

To reach the target temperature in each replica, we used Anderson thermostats with massive collisions after every swap cycle, similar to the approach in refs 51–53 and 59. The choice of the Anderson thermostat is motivated by the fact that, unlike the Berendsen scheme, it rigorously attains a limiting stationary distribution of states that is canonical, as shown in the original derivation in ref 60. In a REMD setting, each between-swap MD trajectory, starting with Boltzmann-randomized initial velocities, is equivalent to a hybrid Monte Carlo move with 100% acceptance (due to energy conservation);⁶¹ as a result, the Markovian evolution due to both swaps and the Andersen thermostats preserves a limiting canonical distribution in each replica and maintains detailed balance. To verify that the swap/collision frequency does not affect our results, we performed comparison simulations for the EK system with a 1 ps frequency and found no qualitative differences in the results from the 20 ps case. In particular, the convergence times and long-time equilibrium ensembles were the same.

Interatomic interactions were modeled using the AMBER ff96 force field⁶² with the implicit solvation model denoted “igb = 5” of Onufriev, Bashford, and Case.⁶³ In a previous study, we and our co-workers found that this force field stabilized well the folds of a number of small peptides⁵² and, when combined with a mechanism-based conformational search algorithm, folded several small proteins.^{51,53} Some earlier efforts found in short simulations that the ff96 force field did not adequately stabilize helical conformations,^{8,39,64,65} although a few groups found it to exhibit α/β balance^{19,21} and perform reasonably well with implicit GBSA models.⁴⁶ We emphasize that, although we show later the performance of this force field is reasonable, provided simulations are carried out long enough, our study here is not primarily a force field test. Rather, our main focus is the general features of convergence in REMD simulations. In addition to the ff96 force field, we also applied the fix proposed by Simmerling and co-workers⁴⁸ that diminishes the intrinsic hydrogen radii on basic residues, in order to bring implicit solvation ion-pairing stabilities in better line with those found in atomistic simulations.

In our initial studies, we considered two variants of the REMD approach. In the first, each cycle, corresponding to

Table 1. Peptides Studied in This Work

peptide/ PDB code	number of residues	secondary structure	disulfide bonds	number of replicas	REMD run length (ns)
1GM2	11	beta	CYS2-CYS10	15	60
C peptide	13	alpha		20	20
EK peptide	14	alpha		20	100 × 5
1O8Y	14	beta	CYS11-CYS3	20	40
15-beta	15	beta		20	20
1E0Q	17	beta		20	20
2I9M	17	alpha		20	80
2JN1	21	beta	CYS3-CYS20	20	40
2JSB	21	beta	CYS3-CYS20	20	40
1WN8	22	alpha		20	60
1B5N	23	beta	CYS7-CYS19	25	100
2EQH	23	beta	CYS1-CYS19	25	200
2DJ9	23	beta	CYS7-CYS19	25	200
1LFC	25	beta	CYS3-CYS20	25	200

one round of MD in each replica followed by swaps, entailed a number of swap attempts equal to the number of temperature neighbors. That is, each possible swap between temperature neighbors was attempted once. In practice, a list of these possible swaps was generated, randomized in order, and then considered in serial succession until the list was exhausted. In a second approach, we increased the number of swaps per temperature neighbor pair per cycle to five. That is, we repeated the aforementioned procedure five times before a second cycle commenced. Since these additional swaps are performed using the same Boltzmann-like acceptance criterion as in the usual REMD approach, it continues to preserve the correct stationary distribution at each temperature.

Simulations of each peptide were analyzed along 5 ns sliding windows across the trajectories; averages corresponding to the window midpoint were computed as a function of time for several quantities. Namely, we computed the rmsd of each peptide backbone from the native or presumed native structure and the population of the most dominant structure after configurational clustering. Clustering was performed using a modified k-means algorithm and a cluster member cutoff of 2 Å rmsd. Populations were determined from the fraction of total trajectory frames that were found to be a member of each cluster, and the structure representing a cluster was taken from the member closest to the cluster centroid. In some cases, we also analyzed the population of intrabackbone hydrogen bonds, for which we used the DSSP criterion⁶⁶ to define the hydrogen-bonding conditions.

Results and Discussion

Folding Heterogeneity in the EK Peptide. Our initial efforts sought to characterize in detail the REMD folding behavior of one peptide, the 14-residue helical EK peptide.⁵⁵ For this peptide, we performed five independent 100 ns runs, each starting from an extended conformation but with different initial Boltzmann-distributed atomic velocities. All of the analyses discussed below pertain to the lowest-temperature replica at 270 K. Times discussed are on a per replica basis, not aggregate over all replicas.

The most conspicuous, and perhaps surprising, observations in these results is that, even for this small a peptide, the folding runs reveal a very high degree of heterogeneity. Figure 1 shows the rmsd (with respect to an ideal helix of

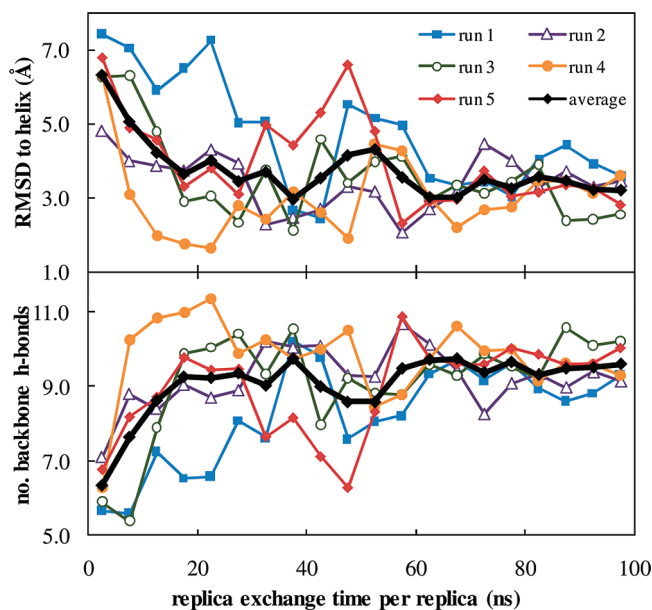


Figure 1. REMD folding of the EK peptide from an extended conformation. (Top) rmsd to an ideal helix as a function of simulation time for five independent runs and their average. (Bottom) Corresponding average number of backbone hydrogen bonds. Each point represents an average over a window of 5 ns centered around it.

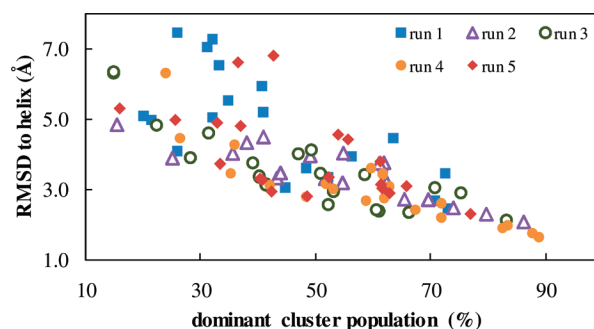


Figure 2. Relationship between rmsd and dominant cluster population for the EK peptide. Each point corresponds to a 5 ns sliding window during the simulation trajectory: the x-axis contains the population of the largest conformational cluster extracted from snapshots during that window, and the y-axis contains the average rmsd for the same snapshots.

identical length) and the hydrogen-bond population as a function of time in the simulation. Run one remains largely unfolded (rmsd > 4.0 Å) for at least the first 30 ns of the simulation, whereas run four rapidly approaches the native state and achieves an rmsd of 2.0 Å in roughly 10 ns. Run five, on the other hand, initially appears to fold at an average rate but experiences a large increase in rmsd at 30 ns that persists for ~20 ns before the system returns to its final average value. An analysis of the run five trajectory reveals that the system folds to a metastable structure during the intermediate times, a β -hairpin structure (shown in Figure 2). In this run, the hairpin forms in higher-temperature replicas and then swaps to the lowest before the system can locate the more favorable helical structure. Overall, despite the small size of this peptide, the collection of runs manifests fast-folding, slow-folding, and misfolding trajectories.

Though the runs differ widely in their individual folding, their mean behavior exhibits some notable trends. Averaged over all five runs, the rmsd versus time curve is well described by a first-order process and exponential decay with a time constant of ~ 8 ns. There is a very strong correlation between rmsd and hydrogen bonding, with low rmsd values almost always paralleling the number of backbone hydrogen bonds formed, in both the average and five individual runs. This result is quite expected for a peptide of this size with a helical structure and geometry which maximizes the amount of intrapeptide hydrogen bonds.

Moreover, we find that convergence in each of the runs can be signaled by a conformational clustering analysis, an indicator that is independent of the native structure. We divided each run into 5 ns windows and performed a clustering of the corresponding trajectory frames; the population of the top cluster was taken as an indicator of the degree of conformational stability within each window. As shown in Figure 2, these dominant cluster populations exhibit a strong correlation with the respective average rmsd of the windows. That is, parts of the simulation that experience less fluctuations appear to be closer to the native structure. Remarkably, the results from the five runs all fall on the same line at low rmsd values and high cluster populations, despite heterogeneities among them. These results suggest the exciting prospect that convergence in each run can be assessed without knowing the native structure but by examining conformational populations. In all runs, a top cluster population of greater than 70% indicates a structure that is of typical rmsd in the equilibrated portions of the simulations. This finding supports the earlier arguments of Daura and co-workers^{34,35} and Lyman and Zuckerman,³⁶ that clustering provides a convenient convergence signal.

How do the peptides reach a folded state in each run? To quantify this behavior, we define a folding event as a point in time at which a folded structure first appears in any of the replicas, given that no folded structure exists in any replica for the previous 100 ps of REMD time. Here, we define “folded” as a snapshot that has an rmsd of 2 Å or less from the ideal helical structure. The distribution of such folding events over the replica temperatures and for the different runs is shown in Figure 3. For each of the five runs, folding events happen in replicas spanning the lowest temperature (270 K) to slightly above the folding temperature ($T_f \sim 350$ K, estimated from the temperature-dependence of the radius of gyration). Moreover, we find that multiple folding events occur in each run due to the fact that folded structures do not always persist through further time evolution. It is interesting to note that the fast folding trajectory (run four) has a higher frequency of folding events near and just below the folding temperature, relative to low temperatures. More strikingly, the misfolding trajectory (run four) has a pronounced peak in folding events near T_f . Both results might suggest that the conformational transitions in these systems stem from higher-temperature events that then propagate to the lowest-temperature replica via swaps. The remaining runs all tend to have higher probabilities of folding events at the lower-temperature replicas. Overall, however,

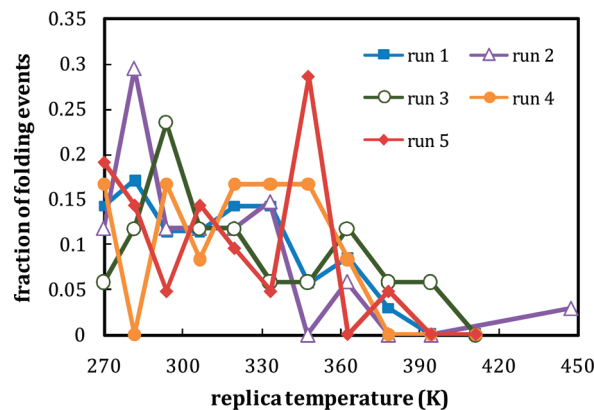


Figure 3. Distribution of folding events in replica space for each EK peptide run. A folding event is defined as a point in time at which a ≤ 2 Å rmsd structure is first reached in any replica given that none exists for the previous 100 ps of REMD time.

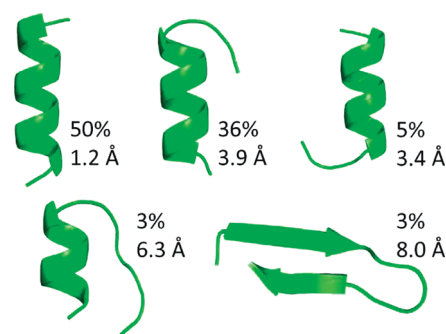


Figure 4. Dominant conformational clusters for the EK peptide. Structures are produced by conformational clustering of the final 40 ns of each of the five independent runs, for a total of 200 ns REMD time. Percentages correspond to the fraction of trajectory snapshots (spaced every 1 ps apart) contained within each cluster group. Rmsd numbers are taken with respect to an energy-minimized ideal helix.

the distributions for all of the runs demonstrate heterogeneity in that folding occurs at different temperatures.

Even after each of the runs converges to its limiting behavior within the 100 ns time frame, the data show that significant conformational fluctuations remain. All five runs appear to have reached a steady-state by a time of ~ 60 ns, beyond where their rmsd values fluctuate around an average of 3.2 Å. Figure 4 shows the structures sampled during these fluctuations and extracted using a clustering analysis of the aggregate final 40 ns from each of the runs (for a total of 200 ns). After equilibration, the peptide spends half of its time near a structure that is very close to the reference ideal helix, at 1.2 Å rmsd. At other times, this system fluctuates to other structures in which either the C-terminal (39% of the time) or N-terminal (5% of the time) parts of the chain partially unfold. In these cases, while the peptide still contains significant helical content, its rmsd from the ideal structure is rather high at 3.4–6.4 Å. The clustering analysis also shows that a very small fraction (1%) of fluctuations locate the β -hairpin structure that was found to be in a misfolded state in the run five trajectory. However, the majority of the conformational fluctuations are centered around a helical

structure, and it is these that contribute to the equilibrium average rmsd of 3.3 Å.

Are these conformational fluctuations representative of the true behavior of the EK peptide found in experiments? Circular dichroism measurements on this peptide reveal only partial helicity near 270 K, which is around a level of 40% when interpreted with a simple Zimm–Bragg model.⁵⁵ It is challenging to determine whether these experiments indicate that the peptide fluctuates between fully and partially folded conformational states or folds to a single, partially helical structure. However, experiments do show the lack of a sharp unfolding transition as urea is added to the EK system, particularly relative to longer helical peptides; this peptide's helicity smoothly and continuously decreases with an increasing urea concentration. Such a low degree of folding cooperativity might suggest that the folded state is not a single structure but evolves from an ensemble of conformations. Thus, it seems reasonable that a peptide of this short size would manifest some degree of structural variety in solution. Moreover, our computed average rmsd values are consistent with those typically found in simulations of short peptides.^{12,17,27,30,33,36} A direct comparison with experiment remains difficult, owing to challenges in bridging the circular dichroism measurements with simulation observables.

It is worthwhile to comment on the role of the force field and the implicit solvation model in these results. Earlier studies with this force field, AMBER ff96, suggested that it overstabilized β -structures,^{8,39,64,65} although a few later investigations by us^{52,53} and others^{19,21,51} found it to exhibit balance between helix and sheet. One potential factor here may be our use of an implicit solvation model, which may favor α -helical structures relative to explicit water simulations via its stabilizing effect on backbone hydrogen bonding.⁴⁹ In that sense, it is possible that a β -propensity in the force field is being corrected fortuitously by an α -propensity in the solvation model.

Alternatively, this work shows that the times scales required for proper folding and convergence greatly exceed those that were attained in earlier studies. Moreover, we find that investigations based on the results of single runs are potentially suspect; in the run five trajectory, for example, a short run would have suggested a strong β -propensity in the force field due to the intermediate formation of a misfolded hairpin structure, when ultimately this state represents only 1% of the equilibrium conformational ensemble. Thus, the possibility remains that early REMD folding studies might have been biased by short simulation times.

Relative to explicit water simulations, the use of an implicit solvation model can contribute to both a speeding up and slowing down of convergence in our runs. On one hand, the lack of any viscosity facilitates a faster exploration of conformational space, and under the premise that an implicit model maintains the same folding free energy, one would expect convergence to be obtained significantly faster with an implicit rather than explicit solvation approach. On the other hand, implicit models have been found to overstabilize ion-pairing interactions, which can lead to the formation of kinetic trap conformations driven by nonnative salt bridges.^{46–49} The EK peptide contains four potential salt bridges, three of which

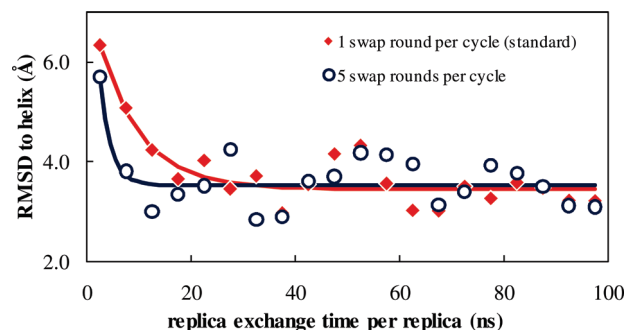


Figure 5. Effect of increasing the number of swaps per round. Each of the two sets of points represents the average over five independent trajectories. The lines correspond to exponential fits of the form $\text{rmsd} = C_1 + C_2 \exp(-t/\tau)$. The red diamonds are identical to the average results reported in the top panel of Figure 1.

can be formed in the native structure. Our analysis shows that, beyond the initial 10 ns of the trajectories, the peptide has, on average, 0.4 salt bridges. We do not find a significant correlation between salt bridge population and convergence behavior, and given the relatively small value of this number, we conclude that these interactions are not slowing convergence for this particular case.

We investigated the possibility of accelerating convergence by increasing the number of swap attempts performed per each REMD cycle. Swaps generally require little computational overhead relative to the molecular dynamics trajectories, and thus, this strategy represents a simple and efficient way to improve performance. In principle, the number of swaps per cycle should have no limiting effect on the peptide configurational ensemble, as these preserve the detailed balance equation, but increasing their number might accelerate convergence by permitting higher-temperature replicas that correctly fold to move more quickly to lower ones.

Figure 5 compares the base case (one swap per neighbor pair per cycle) to the average behavior of five runs in which we increase the number of swap attempts 5-fold. The increase in swaps clearly accelerates mean folding; a detailed analysis shows that the individual runs become less heterogeneous, with the swap increase apparently filtering out slow- and misfolding REMD trajectories. In fitting this new case with an exponential, the time constant is nearly a factor of 3 less than that of the original REMD runs at 3 vs 8 ns. Thus, increasing the number of swaps appears to be a promising strategy for improving REMD convergence by removing slow-folding trajectories with minimal additional overhead.

Convergence for Other Peptides. To broaden our understanding of convergence times beyond the single EK peptide system, we performed REMD folding simulations of 13 additional peptides of 11–25 residues, as shown in Figures 6–10. Here, we ran only one REMD simulation for each peptide, not pursuing multiple runs owing to the size of the data set. For the same reason, we did not perform follow-up runs with an increased number of swap attempts per cycle. However, encouraged by the results we found for the EK peptide, we monitored the dominant cluster population during the progression of each simulation as an indicator of convergence, which is computed in 2 ns sliding windows.

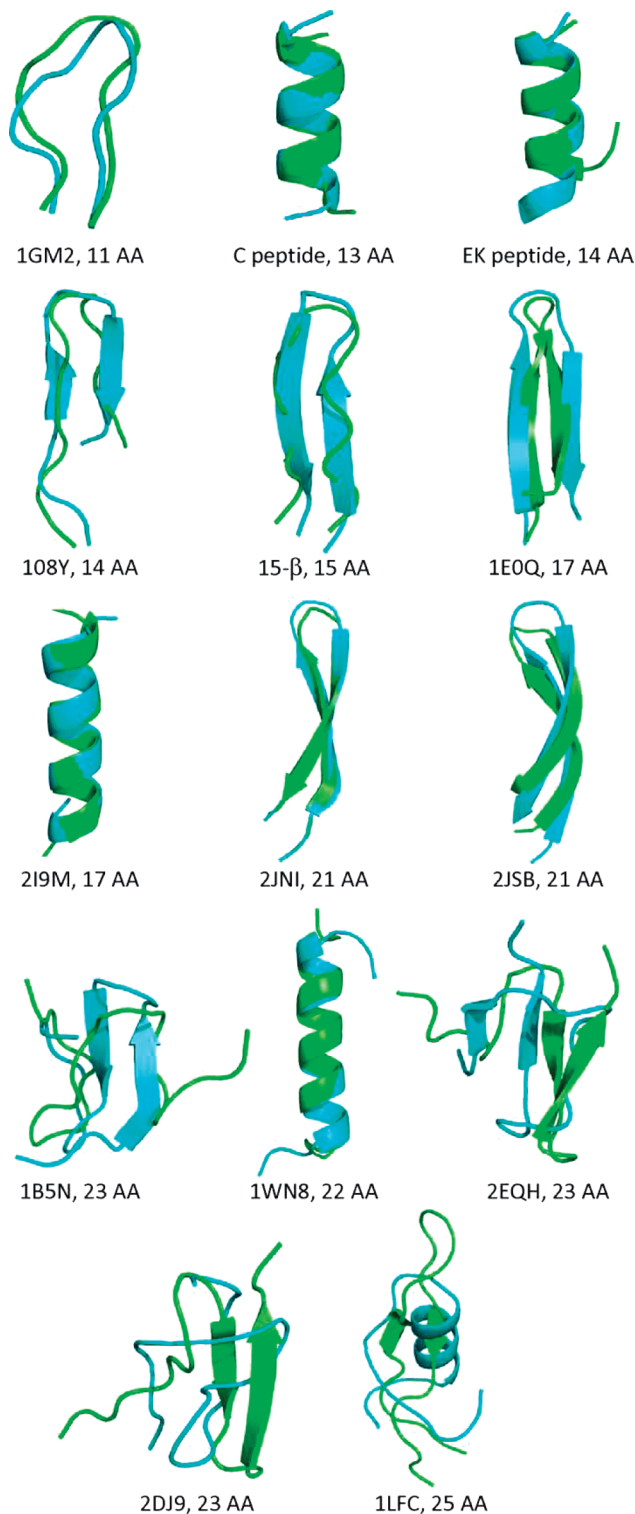


Figure 6. Native (green) versus simulated (blue) structures for peptides folded in this work. For the simulation structures, the dominant conformer from a clustering analysis of the final 10 ns of the simulation run is shown. The number of amino acids is also shown beneath each structure.

Each simulation continued until it had been deemed to reach equilibrium or appeared stable in the native structure, as assessed by the time progression of rmsd. In examining the results of these simulations, our main interests are the degree to which the top cluster population serves as a signal of

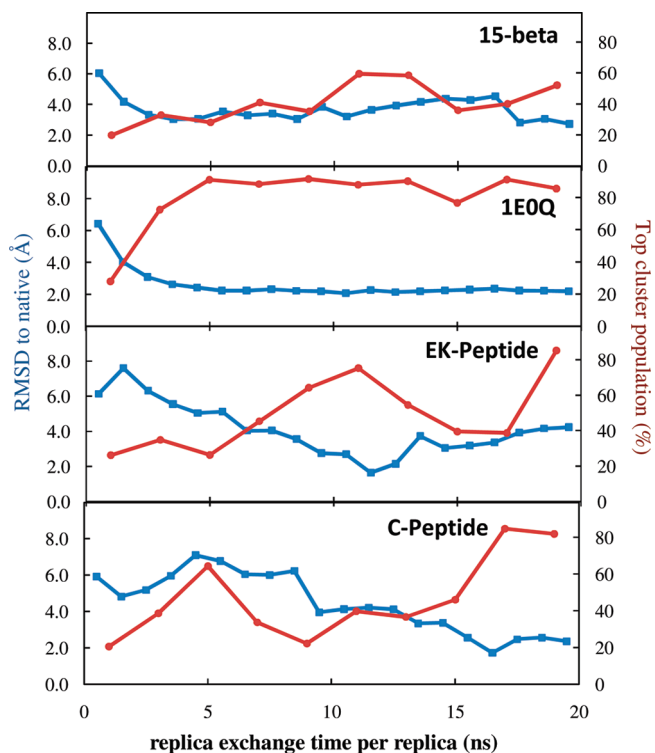


Figure 7. Rmsd and dominant cluster population versus time. Results are shown for the four fastest-folding peptides, whose simulation length was taken to 20 ns. All systems were started from an extended structure.

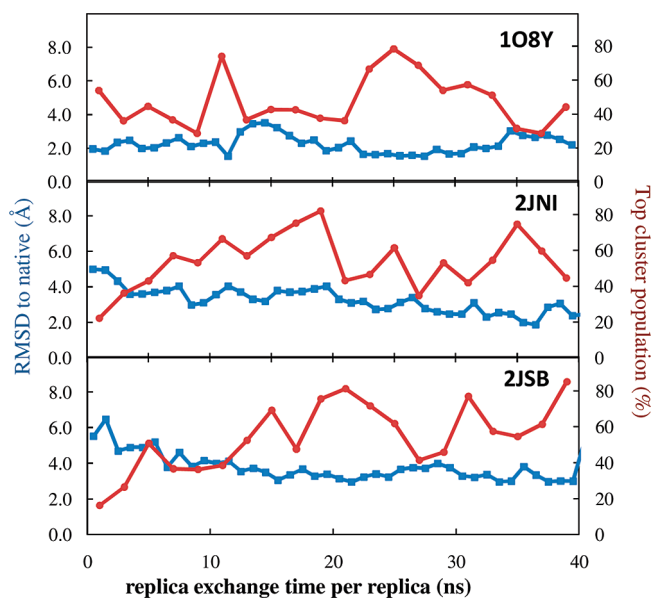


Figure 8. Rmsd and dominant cluster population versus time. Results are shown for peptides whose simulation length was taken to 40 ns. All systems were started from an extended structure.

convergence and the typical time scales involved in simulating to equilibrium this broad sampling of α and β peptides.

Figure 6 shows the top cluster structure from the final 5 ns of each simulation as compared to native, and Figures 7–10 give the rmsd as a function of time for each of the peptides, averaged over 1 ns windows. Of the 14 peptides simulated, 10 ultimately reach average RMSDs that are 4 Å

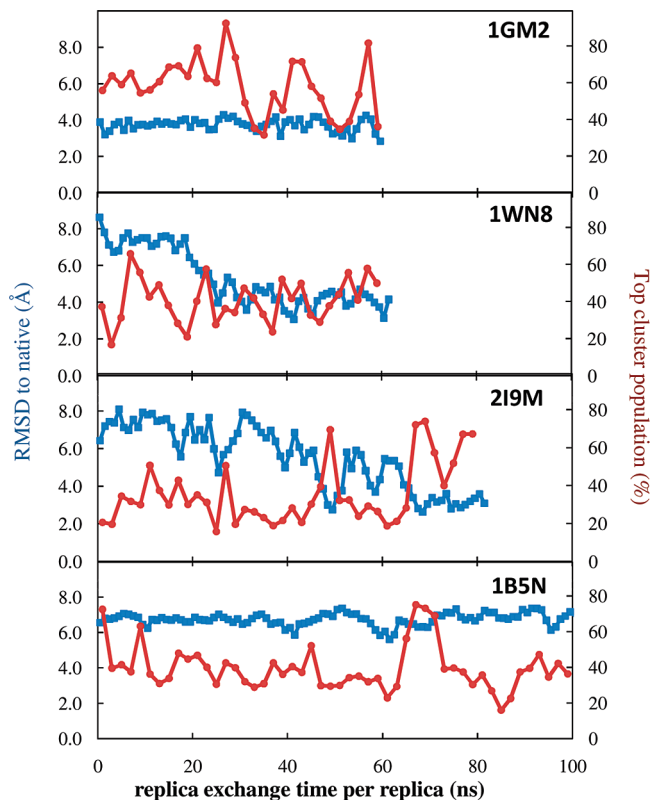


Figure 9. Rmsd and dominant cluster population versus time. Results are shown for peptides whose simulation length was taken to 60–100 ns. All systems were started from an extended structure.

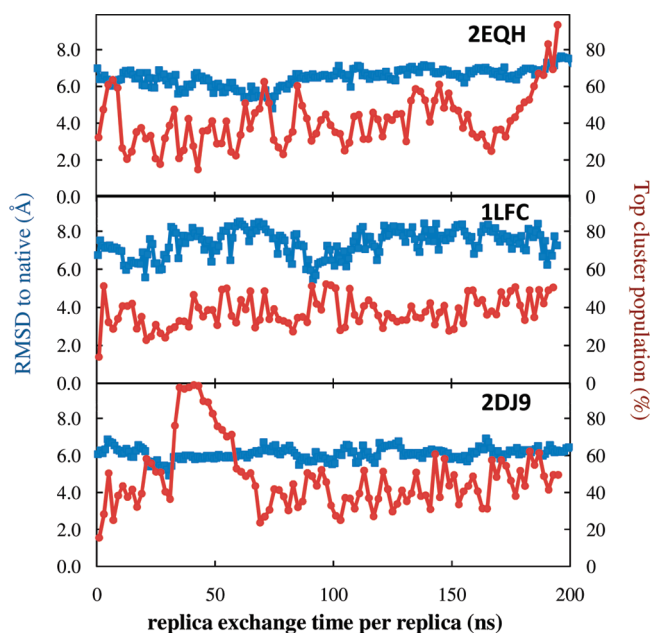


Figure 10. Rmsd and dominant cluster population versus time. Results are shown for peptides whose simulation length was taken to 200 ns. All systems were started from an extended structure.

or less, values that might be considered reasonable for convergence to the native basin given the conformational fluctuations that these short peptides are likely to experience. However, of these 10, there exists a variety of different folding behaviors. The majority of the peptides fold relatively

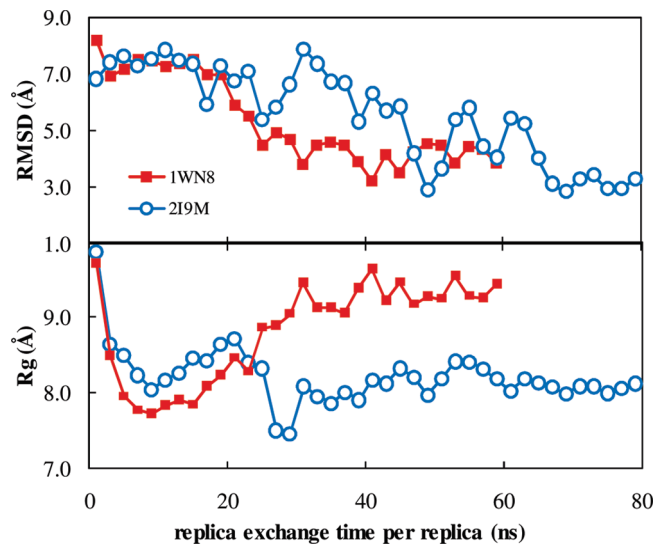


Figure 11. Rmsd and radius of gyration versus time for slow-folding helices. The runs for 1WN8 and 2I9M were terminated at 60 and 80 ns, respectively. Both systems were started from an extended structure.

fast in under 20 ns: the hairpin peptides 15- β , 1E0Q, 1O8Y, 2JNI, 2JSB, and the helical EK and C peptides. It is important to note that three of the hairpin peptides contained high contact order disulfide bonds, which undoubtedly provide a conformational restraint that funnels these simulations toward the native state. Still, 15- β , 1E0Q, and 1O8Y are the fastest folders, converging directly to the native basin in under 5 ns of simulation time, and the former two contain no disulfides at all.

The small 11-residue peptide 1GM2 also appears to converge quickly, its rmsd nearly constant with time from the start of the simulation. Yet, it exhibits a rather high average rmsd from the native structure, fluctuating around a value of 4 Å. The NMR structure for this peptide does contain several conformers, but these at most differ from each other by 1.3 Å, and therefore, the level of structural fluctuations seen in the REMD runs does not reflect the diversity in the PDB data. Still, the top cluster conformation from these runs in Figure 5 is very close to the native structure, a sign that the force field does favor the true native basin, but that significant fluctuations away from this structure raise the time-averaged rmsd. Moreover, it is not unreasonable that a short peptide of this size would exhibit marginal stability as compared to larger ones.

Two helical peptides take appreciably longer to fold. The 22-residue 1WN8 peptide begins to converge toward its native basin just after 20 ns, while the 17-residue 2I9M requires almost 60 ns of REMD simulation. The immediate question arises as to whether or not the slow-folding behavior of these two peptides is due to a β propensity in the backbone secondary structure preferences of the AMBER ff96 force field. In both cases, there is evidence that more specific driving forces play a role. In the case of 1WN8, it appears that a collapse to a dense and broken helix structure, driven by hydrophobic interactions, delays the formation of the full helix. Figure 11 compares the rmsd with the radius of gyration (R_g) versus time; the 1WN8 peptide experiences

an early sharp decrease in R_g that leads to the formation of a compact structure with a few helical turns. Eventually these turns coalesce, and the peptide proceeds to form the full helix with much higher R_g . While helical proteins have been shown to be fast folders,⁶⁷ one distinction here may be the artificial kinetics that REMD and implicit solvation impose.

On the other hand, the slow-folding behavior of the 2I9M helical peptide does not appear to be caused by the drive to a compact shape, as its R_g reaches an equilibrium value well before convergence is evident in the rmsd. We, therefore, clustered snapshots from the intermediate portion of the trajectory, 30–40 ns, and found that the dominant structure was a well-formed β -hairpin. On first look, such evidence might suggest β -propensity in the force field. A closer investigation, however, revealed that ion-pairing interactions between Lys-8, Arg-9, and Glu-12 and Glu-4 and Lys-16 permitted the formation of strong salt bridges that stabilized the turn and terminal hairpin regions well, respectively. Thus, one possible explanation for 2I9M's slow folding might be the presence of kinetic trap conformations dominated by salt bridge interactions; salt bridges have been noted to affect folding in a number of previous studies.^{46–48,52}

The simulations failed to fold four peptides, 1B5N, 2EQH, 1LFC, and 2DJ9. The average rmsd for these four peptides remains near a high constant value for nearly the entire run of 100–200 ns simulation time. All four peptides are long (23 or more residues), contain regions of β structure, have a single high contact order disulfide bond, and have a more complicated structure than that of the simple secondary structure elements of the other peptides. Given the success of the simulations with the other hairpin systems, these results are perhaps a bit surprising. One possibility is that, with the longer length of these peptides, the simulations simply require even more REMD time than currently performed. To assess this possibility, we performed for each a separate “native stability” REMD simulation, in which the initial conformation in all replicas was set to the native structure. Figure 12 shows the rmsd versus time for these runs, where it is clear that, in under 10 ns of simulation time, the system quickly loses the native structure. Therefore, we conclude that the native structure of these four peptides is not stable in the force field used.

Several factors can contribute to the failure of the folding and native stability simulations for these four peptides. One possibility is that the force field may not adequately stabilize larger, compact protein structures that have better-defined hydrophobic cores. This may result from inadequacies in the GBSA solvation model, but may also stem from the fact that a number of terms in the intramolecular force field have been parametrized to calculations on small molecule fragments and folding simulations of short peptides with single secondary structure elements. On the other hand, earlier work by us and others has shown that this force field can reach native-like structures in mechanism-based folding algorithms^{51,53} and can stabilize native folds relative to decoys in “consensus” REMD simulations.⁵³ Another possibility for the failures is that these peptides might be only marginally

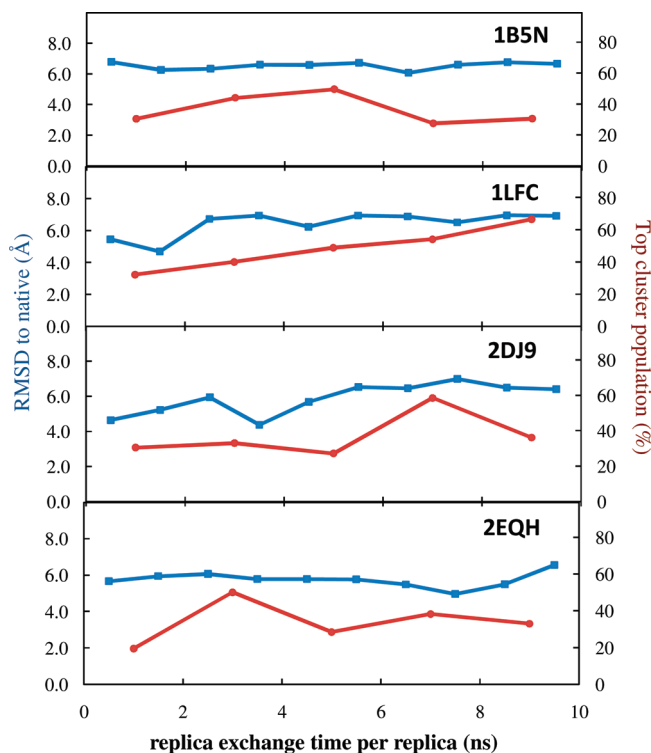


Figure 12. Rmsd and dominant cluster population versus time for “native stability” tests. Shown are the results for peptides that, starting from extended structures, were not able to be folded during 200 ns REMD runs. Here, all systems were started from native PDB structures. The results for each show a loss of structure after less than 2 ns worth of time.

stable in solution at standard conditions. All four have a weak secondary structure content, and all but 1B5N have multiple NMR models that indicate structural diversity in the terminal regions. Interestingly, our simulations of these locate structures with higher secondary structure content including both α and β motifs.

Throughout all of the peptide folding simulations, the dominant cluster population appears to be a fairly consistent predictor of convergence. In eight of the peptide runs (1EOQ, EK, C, 18OY, 2JNI, 2JSB, 1GM2, and 2I9M), a top cluster population of $\sim 70\%$ or more is an indicator of convergence to an ensemble of structures with an average rmsd typical of the equilibrium and folded average rmsd. There are, however, several cases that do not follow this rule. For the 15- β peptide, the criterion must be reduced to the $\sim 60\%$ level, as this peptide never exceeds this threshold population but does correctly fold. For 1WN8, the cluster population reaches 60% at several points during the simulation but does not appear to signal low rmsd values; longer simulation times may show higher populations in this peptide, although we did not pursue them. However, the top cluster criterion also signals some false positives: in three of the peptides that did not fold correctly, it reaches very high values, greater than 90%, at several points during these runs. Part of this behavior may be due to some of the problems with these peptides mentioned earlier, such as their low secondary structure content. Still, this finding does suggest that the

cluster population, while useful, is probably not a foolproof indicator of convergence in all cases.

Conclusions

In this study, we assessed the convergence properties of the REMD method for a number of peptide systems displaying both α and β secondary structure motifs, using an AMBER force field with a GBSA solvation model. We performed a detailed study of the 14-residue helical EK peptide,⁵⁵ conducting five separate 100 ns REMD runs each, starting with an initial unfolded conformation and different velocity set. We then folded 13 additional peptides taken from the PDB and earlier simulation studies. Folding behavior was assessed using backbone rmsd distances from experimental (if available) or idealized (e.g., helical) structures, and a clustering analysis was used to extract dominant conformational states.

Our results suggest several basic features of REMD simulations that are essential to consider in folding studies and, in particular, when using these methods to evaluate force field performance:

Heterogeneity. We find that run-to-run heterogeneity can be quite large. Our multirun study of the EK peptide finds slow-, fast-, and misfolding trajectories that span convergence times from under 10 ns to 30–40 ns. In all runs, folding events are found to occur across a spectrum of temperatures, spanning the lowest replica temperature to near the folding temperature. Moreover, the five different trials for this peptide do not appear to look indistinguishable until reaching a time of about 60 ns. Such heterogeneity is not unexpected given its existence in even fast-folders simulated with molecular dynamics runs,⁵⁰ but it has not yet become common practice to perform multiple trials in REMD folding simulations perhaps due to computational expense. While the degree of heterogeneity observed is likely to be sequence dependent, the possibility of its existence underscores the importance of long and/or multiple simulation trials to achieve a more comprehensive and definitive analysis of folding behavior. One attractive possibility for reducing run-to-run variances is to increase the number of swap attempts performed during REMD cycles; in our study of the EK peptide, we find this approach removes slow-folding trajectories and reduces mean convergence times by a factor of almost three.

Fluctuations. Our simulations demonstrate the presence of significant fluctuations in the converged portions of the folding runs. The average rmsd found for the peptides that do converge to equilibrium and fold to a correct structure varies between 2–4 Å. Typically the dominant cluster structure exhibits an rmsd that is lower than the average, and the nondominant clusters represent fluctuations around it. The average rmsd values that we find are consistent with earlier simulation studies.^{12,17,27,30,33,36} Importantly, these observations reinforce a picture of peptides existing not in static, single structures but in conformational ensemble whose fluctuations might play an important role in their functions in living systems.

Convergence Times. We find that convergence times can be quite large, at least, when compared to the time scales

reached in many earlier efforts. The majority of our peptide folding simulations converge in under 20 ns of REMD time (per replica), but a few systems, notably all helical, require up to 60 ns to reach equilibrium. The factors contributing to these slow-folding times appear to be somewhat distinct for each case, and strong salt bridges seem to contribute to slow-folding in only one of these systems. Importantly, we find a fairly reliable approach for signaling simulation convergence in monitoring the time progression of the population of the largest conformational cluster, which is in agreement with earlier efforts.^{34–36} Importantly, this population metric is independent of the native structure and, thus, suitable for blind, de novo folding studies. For the EK peptide, the dominant cluster population is able to delineate convergence times across the slow- and fast-folding trajectories. For the majority of the other peptides we examined, cluster populations of ~60–70% correspond well with convergence to the native basin. However, this metric is not foolproof; in a few cases, notably those peptides that we were not able to successfully fold, the cluster population gives some false positives for convergence. In future work, the combination of cluster population with other indicators, such as backbone entropies or correlation times, may improve the robustness of this convergence signal.

While these results may be specific to the particular force field that we use, they nevertheless demonstrate the possibility of the kinds of behavior that may accompany any force field and, thus, emphasize the need for addressing convergence issues in force field assessment. Even so, we do find that the combination of AMBER ff96⁶² with OGB implicit solvation⁶³ is able to correctly fold 10 of the 14 peptides studied, a pool including four α -helical and six β -hairpin structures. Its failure to fold the other four peptides may be related to their low secondary structure content, more complicated tertiary arrangement, and unstructured and flexible regions. Overall, this work suggests that the systematic application of long runs, multiple trials, convergence indicators, and diverse peptide databases may offer new perspectives that can improve force field development efforts.

Acknowledgment. We thank the authors in ref 56 for providing the structure of the 15- β peptide. M.S.S. gratefully acknowledges the support of a Camille & Henry Dreyfus Foundation New Faculty Award and a UCSB – Santa Barbara Cottage Hospital Special Research Award.

References

- (1) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical Treatment of Solvation for Molecular Mechanics and Dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.
- (2) Eisenberg, D.; McLachlan, A. D. Solvation energy in protein folding and binding. *Nature* **1986**, *319*, 199–203.
- (3) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141–151.
- (4) Hansmann, U. H. E.; Okamoto, Y. Tertiary Structure Prediction of C-Peptide of Ribonuclease A by Multicanonical Algorithm. *J. Phys. Chem. B* **1998**, *102*, 653–656.
- (5) Schaefer, M.; Bartels, C.; Karplus, M. Solution conformations and thermodynamics of structured peptides: molecular dynam-

- ics simulation with an implicit solvation model. *J. Mol. Biol.* **1998**, *284*, 835–848.
- (6) Dinner, A. R.; Lazaridis, T.; Karplus, M. Understanding beta-hairpin formation. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9068–9073.
- (7) Duan, Y.; Kollman, P. A. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **1998**, *282*, 740–744.
- (8) Ono, S.; Nakajima, N.; Higo, J.; Nakamura, H. Peptide free-energy profile is strongly dependent on the force field: Comparison of C96 and AMBER95. *J. Comput. Chem.* **2000**, *21*, 748–762.
- (9) Ferrara, P.; Caffisch, A. Folding simulations of a three-stranded antiparallel β -sheet peptide. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 10780–10785.
- (10) Jang, S.; Shin, S.; Pak, Y. Molecular Dynamics Study of Peptides in Implicit Water: Ab Initio Folding of Beta-Hairpin, Beta-Sheet, and Beta-Beta-Alpha-motif. *J. Am. Chem. Soc.* **2002**, *124*, 4976–4977.
- (11) Okur, A.; Strockbine, B.; Hornak, V.; Simmerling, C. Using PC clusters to evaluate the transferability of molecular mechanics force fields for proteins. *J. Comput. Chem.* **2003**, *24*, 21–31.
- (12) Simmerling, C.; Strockbine, B.; Roitberg, A. E. All-atom structure prediction and folding simulations of a stable protein. *J. Am. Chem. Soc.* **2002**, *124*, 11258–11259.
- (13) Zagrovic, B.; Sorin, E. J.; Pande, V. Beta-hairpin folding simulations in atomistic detail using an implicit solvent model. *J. Mol. Biol.* **2001**, *313*, 151–169.
- (14) Zagrovic, B.; Snow, C. D.; Shirts, M. R.; Pande, V. S. Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. *J. Mol. Biol.* **2002**, *323*, 927–937.
- (15) Snow, C. D.; Nguyen, H.; Pande, V. S.; Gruebele, M. Absolute comparison of simulated and experimental protein-folding dynamics. *Nature* **2002**, *420*, 102–106.
- (16) Gnanakaran, S.; Nymeyer, H.; Portman, J.; Sanbonmatsu, K. Y.; Garcia, A. E. Peptide folding simulations. *Curr. Opin. Struct. Biol.* **2003**, *13*, 168–174.
- (17) Garcia, A. E.; Sanbonmatsu, K. Y. Exploring the energy landscape of a beta hairpin in explicit solvent. *Proteins* **2001**, *42*, 345–54.
- (18) Zhou, R.; Berne, B. J.; Germain, R. The free energy landscape for beta hairpin folding in explicit water. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 14931–14936.
- (19) Yoda, T.; Sugita, Y.; Okamoto, Y. Secondary structure preferences of force fields for proteins evaluated by generalized-ensemble simulations. *Chem. Phys.* **2004**, *307*, 269–283.
- (20) Pitera, J. W.; Swope, W. Understanding folding and design: Replica-exchange simulations of “Trp-cage” fly miniproteins. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 7587–7592.
- (21) Yoda, T.; Sugita, Y.; Okamoto, Y. Comparisons of force fields for proteins by generalized-ensemble simulations. *Chem. Phys. Lett.* **2004**, *386*, 460–467.
- (22) Anderson, R. E.; Pande, V. S.; Radke, C. J. Dynamic lattice Monte Carlo simulation of a model protein at an oil/water interface. *J. Chem. Phys.* **2000**, *112*, 9167–9185.
- (23) Rao, F.; Caffisch, A. Replica exchange molecular dynamics simulations of reversible folding. *J. Chem. Phys.* **2003**, *119*, 4035–4042.
- (24) Jang, S.; Kim, E.; Pak, Y. Free energy surfaces of miniproteins with a $\beta\beta$ motif: Replica exchange molecular dynamics simulation with an implicit solvation model. *Proteins* **2006**, *62*, 661–671.
- (25) Kim, E.; Jang, S.; Pak, Y. Consistent free energy landscapes and thermodynamic properties of small proteins based on a single all-atom force field employing an implicit solvation. *J. Chem. Phys.* **2007**, *127*, 145104.
- (26) Jang, S.; Kim, E.; Pak, Y. Direct folding simulation of alpha-helices and beta-hairpins based on a single all-atom force field with an implicit solvation model. *Proteins* **2007**, *66*, 53–60.
- (27) Kim, E.; Jang, S.; Pak, Y. Direct folding studies of various α and β strands using replica exchange molecular dynamics simulation. *J. Chem. Phys.* **2008**, *128*, 175104.
- (28) Kim, E.; Yang, C.; Jang, S.; Pak, Y. Free energy landscapes of a highly structured β -hairpin peptide and its single mutant. *J. Chem. Phys.* **2008**, *129*, 165104.
- (29) Wang, J.; Cieplak, P.; Kollman, P. A. How Well Does a Restrained Electrostatic Potential (RESP) Model Perform in Calculating Conformational Energies of Organic and Biological Molecules. *J. Comput. Chem.* **2000**, *21*, 1049–1074.
- (30) Lei, H.; Wu, C.; Wang, Z. X.; Zhou, Y.; Duan, Y. Folding processes of the B domain of protein A to the native state observed in all-atom ab initio folding simulations. *J. Chem. Phys.* **2008**, *128*, 235105.
- (31) Zhang, W.; Wu, C.; Duan, Y. Convergence of replica exchange molecular dynamics. *J. Chem. Phys.* **2005**, *123*, 154105.
- (32) Periole, X.; Mark, A. E. Convergence and sampling efficiency in replica exchange simulations of peptide folding in explicit solvent. *J. Chem. Phys.* **2007**, *126*, 014903.
- (33) Lei, H.; Wu, C.; Liu, H.; Duan, Y. Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 4925–4930.
- (34) Daura, X.; van Gunsteren, W. F.; Mark, A. E. Folding-unfolding thermodynamics of a heptapeptide from equilibrium simulations. *Proteins* **1999**, *34*, 269–280.
- (35) Smith, L. J.; Daura, X.; van Gunsteren, W. F. Assessing equilibration and convergence in biomolecular simulations. *Proteins* **2002**, *48*, 487–496.
- (36) Lyman, E.; Zuckerman, D. M. Ensemble-based convergence analysis of biomolecular trajectories. *Biophys. J.* **2006**, *91*, 164–172.
- (37) Abraham, M. J.; Gready, J. E. Ensuring Mixing Efficiency of Replica-Exchange Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **2008**, *4*, 1119–1128.
- (38) Gnanakaran, S.; Garcia, A. E. Validation of an all-atom protein force field: from dipeptides to larger peptides. *J. Phys. Chem. B* **2003**, *107*, 12555–12557.
- (39) Mu, Y.; Kosov, D. S.; Stock, G. Conformational dynamics of trialanine in water. 2. Comparison of AMBER, CHARMM, GROMOS, and OPLS force fields to NMR and infrared experiments. *J. Phys. Chem. B* **2003**, *107*, 5064–5073.
- (40) Best, R. B.; Buchete, N. V.; Hummer, G. Are Current Molecular Dynamics Force Fields too Helical? *Biophys. J.* **2008**, *95*, 7–9.
- (41) Sorin, E. J.; Pande, V. S. Exploring the Helix-Coil Transition via All-Atom Equilibrium Ensemble Simulations. *Biophys. J.* **2005**, *88*, 2472–2493.

- (42) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **2006**, *65*, 712–725.
- (43) Hu, H.; Elstner, M.; Hermans, J. Comparison of a QM/MM force field and molecular mechanics force fields in simulations of alanine and glycine “dipeptides” (Ace-Ala-Nme and Ace-Gly-Nme) in water in relation to the problem of modeling the unfolded peptide backbone in solution. *Proteins* **2003**, *50*, 451–463.
- (44) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T. A point-charge force field for molecular mechanics simulations of proteins. *J. Comput. Chem.* **2003**, *24*, 1999–2012.
- (45) Zhou, R.; Berne, B. J. Can a continuum solvent model reproduce the free energy landscape of a β -hairpin folding in water. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 12777–12782.
- (46) Zhou, R. Free energy landscape of protein folding in water: explicit vs. implicit solvent. *Proteins* **2003**, *53*, 148–161.
- (47) Felts, A. K.; Harano, Y.; Gallicchio, E.; Levy, R. M. Free energy surfaces of beta-hairpin and alpha-helical peptides generated by replica exchange molecular dynamics with the AGBNP implicit solvent model. *Proteins* **2004**, *56*, 310–321.
- (48) Geney, R.; Layten, M.; Gomperts, R.; Hornak, V.; Simmerling, C. Investigation of salt bridge stability in a generalized Born solvent model. *J. Chem. Theory Comput.* **2006**, *2*, 115–127.
- (49) Roe, D. R.; Okur, A.; Wickstrom, L.; Hornak, V.; Simmerling, C. Secondary structure bias in generalized Born solvent models: comparison of conformational ensembles and free energy of solvent polarization from explicit and implicit solvation. *J. Phys. Chem. B* **2007**, *111*, 1846–1857.
- (50) Ensign, D. L.; Kasson, P. M.; Pande, V. S. Heterogeneity Even at the Speed Limit of Folding: Large-scale Molecular Dynamics Study of a Fast-folding Variant of the Villin Headpiece. *J. Mol. Biol.* **2007**, *374*, 806–816.
- (51) Ozkan, S. B.; Wu, G. H. A.; Chodera, J. D.; Dill, K. A. Protein folding by zipping and assembly. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 11987–11992.
- (52) Shell, M. S.; Ritterson, R.; Dill, K. A. A test on peptide stability of AMBER force fields with implicit solvation. *J. Phys. Chem. B* **2008**, *112*, 6878–6886.
- (53) Shell, M. S.; Ozkan, S. B.; Voelz, V. A.; Wu, G. H. A.; Dill, K. Blind test of physics-based prediction of protein structures. *Biophys. J.* **2009**, *96*, 917–924.
- (54) Murzin, A. G.; Brenner, S. E.; Hubbard, T.; Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **1995**, *247*, 536–540.
- (55) Scholtz, J. M.; Barrick, D.; York, E. J.; Stewart, J. M.; Baldwin, R. L. Urea Unfolding of Peptide Helices as a Model for Interpreting Protein Unfolding. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 185–189.
- (56) Santiveri, C. M.; Pantoja-Uceda, D.; Rico, M.; Jimenez, M. A. Beta-hairpin formation in aqueous solution and in the presence of trifluoroethanol: a (^1H) and (^{13}C) nuclear magnetic resonance conformational study of designed peptides. *Biopolymers* **2005**, *79*, 150–162.
- (57) Bierzynski, A.; Kim, P. S.; Baldwin, R. L. A Salt Bridge Stabilizes the Helix Formed by Isolated C-Peptide of RNase A. *Proc. Natl. Acad. Sci. U.S.A.* **1982**, *79*, 2470–2474.
- (58) Pitera, J. W.; Haque, I.; Swope, W. C. Absence of reptation in the high-temperature folding of the trpzip2 β -hairpin peptide. *J. Chem. Phys.* **2006**, *124*, 141102.
- (59) Ho, B. K.; Dill, K. A. Folding very short peptides using molecular dynamics. *PLoS Comput. Biol.* **2006**, *2*, e27.
- (60) Andersen, H. C. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.* **1980**, *72*, 2384.
- (61) Mehlig, B.; Heermann, D. W.; Forrest, B. M. Hybrid Monte Carlo method for condensed-matter systems. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1992**, *45*, 679.
- (62) Kollman, P. A.; Dixon, R.; Cornell, W.; Fox, T.; Chipot, C.; Pohorille, A. The development/application of a “minimalist” organic/biochemical molecular mechanic force field using a combination of ab initio calculations and experimental data. *Comput. Simul. Biomol. Syst.* **1997**, *3*, 83–96.
- (63) Onufriev, A.; Bashford, D.; Case, D. A. Modification of the Generalized Born model suitable for macromolecules. *J. Phys. Chem. B* **2000**, *104*, 3261–3429.
- (64) García, A. E.; Sanbonmatsu, K. Y. alpha-Helical stabilization by side chain shielding of backbone hydrogen bonds. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 2782–2787.
- (65) Gnanakaran, S.; Garcia, A. E. Folding of a highly conserved diverging turn motif from the SH3 domain. *Biophys. J.* **2003**, *84*, 1548–1562.
- (66) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577–2637.
- (67) Plaxco, K. W.; Simons, K. T.; Baker, D. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **1998**, *277*, 985–994.

CT900119N

JCTC Journal of Chemical Theory and Computation

Absorption Spectrum of the Green Fluorescent Protein Chromophore: A Difficult Case for *ab Initio* Methods?

Claudia Filippi,^{*,†,‡,§} Maurizio Zaccheddu,[†] and Francesco Buda^{*,§}

Instituut-Lorentz, Universiteit Leiden, Niels Bohrweg 2, 2333 CA Leiden, The Netherlands, Faculty of Science and Technology and MESA+ Research Institute, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands, and Leiden Institute of Chemistry, P.O. Box 9502, 2300 RA Leiden, The Netherlands

Received May 7, 2009

Abstract: We perform a thorough comparative investigation of the excitation energies of the anionic and neutral forms of the green fluorescent protein (GFP) chromophore in the gas phase using a variety of first-principle theoretical approaches commonly used to access excited state properties of photoactive molecules. These include time-dependent density functional theory (TDDFT), complete-active-space second-order perturbation theory (CASPT2), equation-of-motion coupled cluster (EOM-CC), and quantum Monte Carlo (QMC) methods. We find that all approaches give roughly the same vertical excitation for the anionic form, while TDDFT predicts an excitation for the neutral chromophore significantly lower than the highly correlated methods. Our findings support the picture emerging from the extrapolation of the Kamlet-Taft fit of absorption experimental data in solution and indicate that the protein gives rise to a considerable bathochromic shift with respect to vacuum. These results also open some questions on the interpretation of photodestruction spectroscopy experiments in the gas phase as well as on the accuracy of previous theoretical calculations in the more complex protein environment.

1. Introduction

Computational modeling is a crucial complement to experiments in deepening our understanding of optical processes occurring in photosensitive biological systems. Despite significant theoretical progress in electronic structure methods, it is far from trivial to accurately compute excitation properties of even relatively small photoactive molecules. It is not unusual in surveying the vast theoretical literature on photosensitive biosystems to find that the large spread of first-principle approaches used for a particular case yields an equally large spread of results and predictions.

The difficulties arise from the rather strict requirements the theoretical approach must meet to provide a predictive description of a photoactive biosystem. It should give an

accurate and balanced description of the ground and the excited states of the photoactive site and also be able to treat a realistically large model of the biosystem. Time-dependent density functional theory (TDDFT)¹ is the most appealing approach to efficiently compute the excitations of large molecular complexes. However, known shortcomings of conventional adiabatic TDDFT as the underestimation of charge transfer excitations^{2–4} or the lack of explicit inclusion of two- and higher-electron excitations^{5–8} may severely affect the accuracy in the case of photoactive molecules. Therefore, despite their less favorable scaling with system size, highly correlated quantum chemical methods such as complete-active-space second-order perturbation theory (CASPT2)⁹ have established themselves as the main theoretical tool to compute excitations of photosensitive systems^{10,11} even for relatively large systems where compromises must be taken in the choice of the size of the active space or of the basis set.

The green fluorescent protein (GFP) first isolated in the jellyfish *Aequorea victoria*¹² is an intrinsically fluorescent

* Corresponding author e-mail: filippi@lorentz.leidenuniv.nl (C.F.), f.buda@chem.leidenuniv.nl (F.B.).

[†] Instituut-Lorentz Leiden.

[‡] Faculty of Science and Technology Twente.

[§] Leiden Institute of Chemistry.

protein which has revolutionized cellular biology in the last decades.^{13–15} Its most numerous and successful applications are as a genetic fusion partner to host proteins which maintain their normal functions but are then fluorescent and can be dynamically visualized in living cells and organisms. Beyond being extremely relevant in biotechnology, GFP represents a perfect playground for theoretical investigation of photoactive biomolecules due to several reasons. First, the spectroscopic features of the chromophore of GFP have been experimentally extensively studied in vacuo,^{16–18} solution,^{19–24} and in the protein environment,^{25–30} offering a wealth of data for theoretical comparison. Then, GFP has already been the subject of a large number of computational semiempirical^{31–33} and first-principle^{34–46} studies, which often differ in the conclusions and do not therefore offer a consistent picture. For instance, early semiempirical calculations incorrectly assigned the charge state of the chromophore to the absorption bands in the protein.³¹ On the other hand, first-principle TDDFT^{42–44} and CASPT2³⁷ calculations significantly differ in the prediction of the absorption spectrum in vacuo, while these methods in combination with different quantum-mechanics/molecular-mechanics (QM/MM) schemes^{36,38} yield results in agreement with each other and with the experimental spectrum in the more complex protein environment. These findings have opened new questions which require a more thorough comparative investigation already at the level of simple gas-phase models.

In this paper, we employ a variety of first-principle theoretical approaches to compute the electronic excitations in the gas phase of the neutral and anionic forms of the GFP chromophore. These are the protonation states relevant in the protein and responsible for the two room-temperature absorption peaks of wild-type GFP at 398 nm (3.12 eV) and 478 nm (2.59 eV), respectively.²⁹ In addition to the neutral and anionic chromophores, we consider a cationic model compound which was recently characterized in gas-phase spectroscopy experiments.¹⁸ For all these models, we compare the performance of TDDFT and highly correlated techniques such as CASPT2, equation-of-motion coupled cluster (EOM-CC),⁴⁷ and the less widely used quantum Monte Carlo (QMC) methods.⁴⁸ We find that the treatment of the neutral form is more problematic for all theoretical approaches due to a stronger multiconfigurational nature of its electronic states as compared to the anionic case where a small complete-active-space (CAS) wave function is sufficient to converge both QMC and CASPT2 calculations. For the neutral chromophore, TDDFT predicts a vertical excitation significantly lower than the other correlated approaches, while all theoretical methods yield practically the same excitation energy for the anionic form. We note that a multistate approach has to be adopted in the CASPT2 calculations for the neutral chromophore and that our CASPT2 results for the anionic form are at variance with previous studies³⁷ using the same technique but a different definition of the zero-order Hamiltonian. Our findings are rather puzzling and raise concerns on the interpretation of photodestruction spectroscopy experiments in the gas phase^{16–18} as well as on the accuracy of previous TDDFT³⁶ and CASPT2³⁸ calculations in the more complex protein

environment. Differently from what had been inferred from these experiments and calculations, our theoretical excitations in the gas phase indicate that the bathochromic shifts in the protein are quite significant. Our results are consistent with values extrapolated from absorption experiments in solution.²²

In Section 2, we briefly present the methods used in this paper and focus on the description of a novel QMC scheme where we simultaneously optimize orthogonal Jastrow-Slater wave functions for ground and excited states. Computational details are given in Section 3. The model chromophores are described in Section 4, and the results are presented in Section 5. Finally, discussion and conclusions are in Section 6.

2. Methods

In this paper, we employ a variety of first-principle quantum chemical methods. We omit here the description of the more traditional TDDFT, CASSCF, CASPT2, and CC approaches since these methods have become a well-established part of quantum chemistry, and their description can be found in appropriate textbooks.^{1,49} In contrast, QMC methods⁴⁸ represent a less explored alternative to conventional highly correlated approaches and as such deserve a brief presentation. Moreover, the method used to optimize excited state QMC wave functions is novel and is therefore described in detail below.

2.1. QMC and State-Average Optimization of Multiple States. QMC methods provide an efficient and accurate description of both dynamical and static electronic correlation and have been mostly employed for ground-state studies of molecular and extended systems.⁴⁸ Their application to the computation of excited states of small photosensitive molecules is more recent and appears very encouraging.^{50–53}

The key ingredient which determines the quality of a QMC calculation is the many-body trial wave function which, in the present work, is chosen of the Jastrow-Slater type. As we treat multiple states of the same symmetry, we write the ground- and excited-state wave functions as linear combination of spin-adapted configuration state functions (CSF) multiplied by a Jastrow correlation factor

$$\Psi^I = \sum_{i=1}^{N_{\text{CSF}}} c_i^I \mathcal{J} \mathcal{C}_i \quad (1)$$

where different states depend on their individual linear coefficients c_i^I but share a common set of single-particle orbitals and Jastrow factor \mathcal{J} . We use here a Jastrow factor which correlates pairs of electrons and each electron separately with a nucleus and employ different Jastrow factors to describe the correlation with different atom types. Since the optimal orbitals and expansion coefficients in Ψ^I may differ from the values obtained for instance in a CASSCF calculation in the absence of the Jastrow factor \mathcal{J} , it is important to reoptimize them in the presence of the Jastrow component.

Here, we present an efficient and simple approach to obtain accurate and orthogonal many-body wave functions for multiple states of the same symmetry.^{52,53} We obtain the

Jastrow and orbital parameters which minimize the average energy over the state of interest and the lower states, while the linear coefficients in the CSF expansion ensure that orthogonality is preserved among the states. In the orbital and Jastrow optimization step, we extend the linear optimization method for ground states⁵⁴ to the optimization of multiple states in a state average fashion. The resulting scheme is simpler and superior to the approach we previously proposed⁵⁰ where only the orbitals were optimized and orthogonality was only approximately preserved among the states.

It is well-known that an optimal set of linear coefficients is readily obtained by solving the generalized eigenvalue problem

$$\sum_{j=1}^{N_{\text{CSF}}} H_{ij} c_j^I = E_I \sum_{j=1}^{N_{\text{CSF}}} S_{ij} c_j^I \quad (2)$$

The Hamiltonian and overlap matrix elements are estimated by a finite-sample average in variational Monte Carlo as

$$H_{ij} = \left\langle \frac{\mathcal{J}C_i \mathcal{H} \mathcal{J}C_j}{\Psi_g} \right\rangle_{\Psi_g^2}, S_{ij} = \left\langle \frac{\mathcal{J}C_i \mathcal{J}C_j}{\Psi_g} \right\rangle_{\Psi_g^2} \quad (3)$$

where the statistical average is over the Monte Carlo configurations sampled from Ψ_g^2 . To obtain an accurate estimate of the optimal coefficients of more than one state, the guiding wave function Ψ_g should be chosen to have significant overlap with all states of interest. In particular, we use here $\sqrt{\sum_l |\Psi_l|^2}$. Importantly, it has been shown that the use of the nonsymmetric estimator of the Hamiltonian matrix of eq 3 yields a strong zero-variance principle and results in a particularly efficient approach.⁵⁵ The scheme has also the advantage to enforce orthogonality between the wave functions of the multiple states and ensure a generalized variational principle.

To optimize the nonlinear parameters in the Jastrow factor and orbitals of the multiple states, we follow a state-average (SA) approach to determine a set of orbitals and a Jastrow factor which give a comparably good description of the states under considerations while preserving orthogonality among the states. We alternate between optimizing the linear coefficients and the nonlinear (Jastrow and orbital) coefficients in which the quantity minimized is the weighted average of the energies of the states under consideration

$$E_{\text{SA}} = \sum_{I \in A} w_I \frac{\langle \Psi^I | \mathcal{H} | \Psi^I \rangle}{\langle \Psi^I | \Psi^I \rangle} \quad (4)$$

where the weights w_I are fixed and $\sum_I w_I = 1$. Therefore, at convergence, the averaged energy E_{SA} is stationary with respect to all parameter variations subject to the orthogonality constraint, while the individual state energies E_I are stationary with respect to variations of the linear coefficients but not with respect to variations of the orbital or Jastrow parameters. In this approach, the wave functions are kept orthogonal and a generalized variational theorem applies.

To improve the orbital and Jastrow parameters at each SA iteration step, we extend the ground-state linear optimization approach⁵⁴ to the SA optimization of multiple states. Under a common variation in an orbital or Jastrow parameter p_i ,

the changes in the states Ψ^I are given by $\Psi_k^I = (\partial \Psi^I / \partial p_k)$ and can be made orthogonal to the corresponding state as

$$\bar{\Psi}_k^I = \Psi_k^I - \left[\left\langle \frac{\Psi_k^I \Psi_k^I}{\Psi_g} \right\rangle_{\Psi_g^2} \left\langle \frac{\Psi_g^2}{\Psi_g} \right\rangle_{\Psi_g^2} \right] \Psi^I \quad (5)$$

To linearize the minimization with respect to the nonlinear parameters, we work in the semiorthogonal basis of the functions $\{\bar{\Psi}_0^I, \bar{\Psi}_k^I\}$, where $\bar{\Psi}_0^I = \Psi^I$, and find the variations Δp_i in the parameters as the lowest eigenvalue solution of the generalized eigenvalue problem

$$H_{ij}^{\text{SA}} \Delta p_j = E_{\text{SA}} \Delta p_j \quad (6)$$

where $\Delta p_0 = 1$. The SA Hamiltonian matrix is computed as

$$H_{ij}^{\text{SA}} = \sum_{I \in A} w_I \left[\left\langle \frac{\bar{\Psi}_i^I \mathcal{H} \bar{\Psi}_j^I}{\Psi_g} \right\rangle_{\Psi_g^2} \left\langle \frac{\Psi_g^2}{\Psi_g} \right\rangle_{\Psi_g^2} \right] \quad (7)$$

and an analogous definition holds for the SA overlap matrix elements. The matrix elements for all states are computed in a single variational Monte Carlo run with a guiding wave function Ψ_g chosen to have significant overlap with the states of interest. At convergence and for the optimal linear coefficients, the minimal energy E_{SA} (eq 4) is obtained: if the iterative scheme converges, the matrix elements H_{ij}^{SA} are zero, and, consequently, the derivatives of E_{SA} with respect to the parameter p_i are zero as they equal H_{i0}^{SA} .

In summary, one iteration of excited state optimization consists of the following steps: *i*) sample the quantities needed for the optimization of the linear coefficients with the appropriate guiding wave function; *ii*) diagonalize the matrix (eq 2) to obtain the optimal linear coefficients for the states under consideration; *iii*) sample for all states the quantities needed in the linear equations (eq 6) and obtain the parameters Δp_i ; and *iv*) construct a set of improved orbitals and Jastrow parameters as $p_i \rightarrow p_i + \Delta p_i$. As in the optimization of a ground state wave function, when the nonlinear parameters are far from the optimal values, the optimization may need to be stabilized by shifting all diagonal elements except the first one as $H_{ij}^{\text{SA}} \rightarrow H_{ij}^{\text{SA}} + a_{\text{diag}} \delta_{ij} (1 - \delta_{i0})$ where a_{diag} is automatically determined in a correlated sampling run as described in ref 54.

The trial wave functions are then used in diffusion Monte Carlo (DMC), which produces the best energy within the fixed-node approximation [i.e., the lowest-energy state with the same zeros (nodes) as the trial wave function].

3. Computational Details

The DFT calculations are carried out using the Gaussian 03⁵⁶ and the Amsterdam Density Functional (ADF)^{57–59} code. The Gaussian code is used to optimize the ground state geometries of the chromophores within all-electron DFT with the BLYP⁶⁰ and B3LYP⁶¹ functionals. The same code is also used to perform the all-electron linear-response adiabatic TDDFT calculations with the BLYP and B3LYP functionals at the corresponding ground state structures. All Gaussian calculations use a cc-pVTZ basis set and default convergence criteria. The ADF code is employed to perform all-electron

TDDFT calculations using the state-average orbital-dependent potential (SAOP)⁶² in the eigenvalue difference matrix of the response equations, in combination with the LDA exchange-correlation potential in the linear response kernel and for the computation of the reference state. The Slater ET-pVQZ basis set is used in all ADF calculations.

The program MOLCAS 7.2⁶³ is used for the all-electron CASPT2 and multistate (MS) CASPT2⁶⁴ calculations. The starting SA-CASSCF wave functions are obtained with equal weights over the states of interest. We employ the default IPEA zero-order Hamiltonian,⁶⁵ and we indicate if an additional constant level shift⁶⁶ is added to the Hamiltonian. Most calculations are done with the cc-pVDZ basis set, but other basis sets such as 6-31G*, 6-31+G*, and cc-pVTZ are also tested. In the CASPT2 calculations, we do not correlate as many lowest orbitals of σ character as the number of heavy atoms in the model.

The EOM-CC calculations are performed with the code MOLPRO 2006.1.⁶⁷ The CC calculations include single and double excitations (CCSD) and employ both the cc-pVDZ basis and a basis we denote with cc-pVTZ' where the cc-pVTZ basis is used for all atoms except hydrogen which is described by a cc-pVDZ basis. We do not correlate as many lowest orbitals of σ character as the number of heavy atoms in the model.

The program package CHAMP⁶⁸ is used for the QMC calculations. We employ scalar-relativistic energy-consistent Hartree-Fock pseudopotentials⁶⁹ where the carbon, nitrogen, and oxygen 1s electrons are replaced by a nonsingular s -nonlocal pseudopotential and the hydrogen potential is softened by removing the Coulomb divergence. We employ the cc-pVDZ Gaussian basis sets⁶⁹ constructed for these pseudopotentials. Different Jastrow factors are used to describe the correlation with a hydrogen, oxygen, and carbon atom. For each atom type, the Jastrow factor consists of an exponential of the sum of two-fifth-order polynomials of the electron-nuclear and the electron-electron distances, respectively.⁷⁰ The starting determinantal components are obtained in CASSCF calculations which are performed with the program GAMESS(US).⁷¹ In all SA-CASSCF calculations, equal weights are employed for the states, and the final CAS expansions are expressed on the weighted-average CASSCF natural orbitals. The CAS wave functions of the ground and excited states may be truncated with an appropriate threshold on the CSF coefficients, and the union set of surviving CSFs for the states of interest is retained in the QMC calculations. The Jastrow correlation factor and the CI coefficients are optimized by energy minimization in a state-averaged sense within VMC as described in the previous Section, where equal weights are used in the optimization. When indicated in the text, also the orbitals are optimized along with the Jastrow and CI parameters. The pseudopotentials are treated beyond the locality approximation,⁷² and imaginary time steps of 0.055 or 0.075 au are used in the DMC calculations.

4. Chromophore Models

The fluorescent chromophore of GFP consists of a p -hydroxybenzylideneimidazolinone⁷³ (p -HBI) which sits in

a β -can structure formed by the protein²⁶ and, in wild-type GFP, exists in two protonation states, neutral and anionic, corresponding to a protonated (1C) and deprotonated (1A) phenolic oxygen, respectively.^{19,25,27,30,32,35} The chromophore models studied in this paper are depicted in Figure 1 and can be divided in three groups: The anionic chromophores (A, B), the neutral chromophores (C, D), and a positively charged chromophore (E). We always refer to Figure 1 and its labels when describing the models in the rest of the paper.

The p -HBI anionic *minimal* model (A) is the smallest possible representation of the GFP chromophore and possesses C_s symmetry. Given its favorable size and symmetry, we study very extensively this model (and its neutral counterpart) using a variety of correlated techniques. This chromophore was also employed in previous correlated CASPT2 calculations,³⁷ but no experimental characterization is available. The anionic *methyl-terminated* model (B) has also C_s symmetry and only differs from the minimal model (A) in the termination with two methyl groups instead of the hydrogens in the imidazolone ring (p -HBDI). Even though we expect its electronic properties to be rather similar to those of model (A), we study this chromophore since it was synthesized and investigated in gas-phase photodestruction spectroscopy experiments, which place its absorption maximum at 2.59 eV,¹⁶ that is, very close to the value of the anionic chromophore in the protein. The p -HBDI (B) model was also studied in absorption experiments in solution.²²

For the neutral chromophore, we construct two models analogous to the anionic case, namely the minimal (C) and the methyl-terminated (D) model. To the best of our knowledge, no experiments are available for these neutral chromophores in the gas phase. The methyl-terminated (D) model was studied in absorption experiments in solution.²²

The experimental photodestruction spectroscopy technique employed to study the anionic chromophore (B) makes use of an electrostatic ion storage ring and can therefore be applied only to charged molecules. Therefore, a positively charged compound (E) was synthesized by adding a methyleneammonium cation to the neutral p -HBDI chromophore (D) to mimic the spectrum of the neutral chromophore.¹⁸ We refer to model (E) as the *neutral*⁺ chromophore. In the photodestruction spectroscopy experiment, the absorption maximum of model (E) is located at 2.99 eV.

5. Results

5.1. Structural Analysis of the Models. The structures of the chromophores are relaxed in the ground state using DFT and various exchange-correlation functionals. We discuss in detail the geometrical features of the minimal anionic (A) and neutral (C) models as the same bond-length patterns are observed for the methyl-terminated models. The neutral⁺ (E) chromophore requires a separate discussion. For the atom labeling, we refer the reader to Figure 2.

The bond lengths of the minimal anionic (A) and neutral (C) models optimized with the BLYP functional and a cc-pVTZ basis are shown in Figure 3. We only show the bond

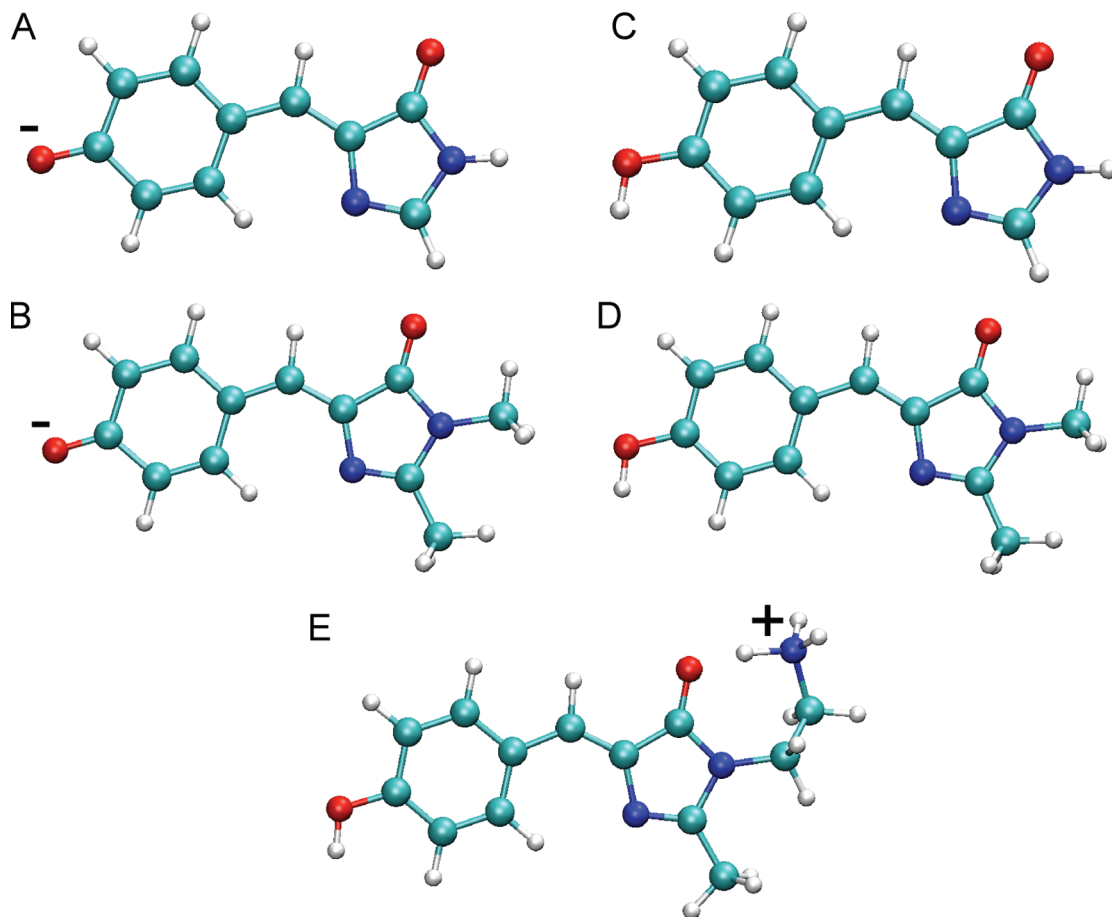


Figure 1. Gas-phase chromophore models: the anionic minimal (A) and methyl-terminated (B) models; the neutral minimal (C) and methyl-terminated (D) models; and the positively charged (neutral⁺) model (E).

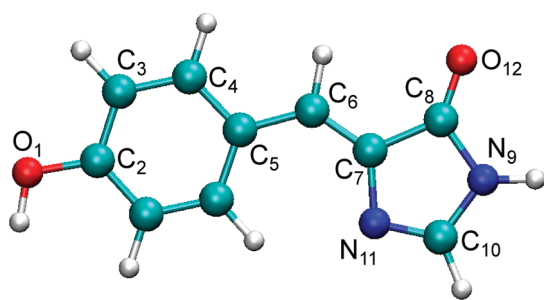


Figure 2. Atom numbering used for the chromophore of GFP.

lengths obtained with this exchange-correlation functional as other functionals yield practically equivalent geometries. In particular, the bond lengths obtained with the hybrid B3LYP functional are always shorter by only about 0.01 Å than the BLYP values.

Since the anionic chromophore may pose some difficulties for DFT, we also perform a CASPT2 optimization of the minimal anionic (A) model using a CAS(12,11) expansion (12 π -electrons distributed in 11 π -orbitals) and the 6-31G* basis set. The CASPT2 bond lengths are shown in Figure 3 where we also report the CASSCF bond lengths obtained in ref 37 using the same basis set and active space. We find that the CASSCF bond lengths are systematically shorter than the BLYP ones with an average deviation smaller than 0.02 Å and that the use of the CASPT2 method yields a geometry in better agreement with the BLYP one reducing the

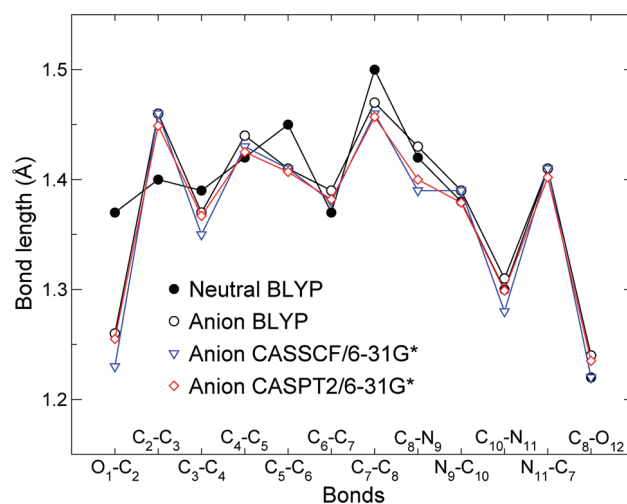


Figure 3. Bond lengths of the minimal anionic (A) and neutral (C) models optimized within DFT/cc-pVTZ with the BLYP functional. For the anionic model, the results from CASSCF/6-31G*³⁷ and CASPT2/6-31G* calculations are also shown. The bonds of the central carbon bridge are C₅-C₆ and C₆-C₇.

deviation to 0.01 Å. Therefore, since also the CASSCF and CASPT2 geometries are very close to the BLYP structure, we can safely claim that, when we compare excitation spectra for gas-phase models, the eventual differences one may observe should be attributed to the theoretical approaches

employed to compute the excitations rather than the approach adopted to optimize the geometry.

While the main structural features are largely independent of the theoretical method, the most evident difference is between the bond lengths of the anionic (A) and the neutral (C) model as shown in Figure 3. The geometrical changes between neutral and anionic chromophore can be rationalized in terms of the resonant benzenoid and quinoid forms which are accessible upon deprotonation of the phenolic oxygen.³³ The neutral model is characterized by the aromaticity of the phenolic ring with rather similar bond lengths and by a single–double bond-length alternation at the carbon bridge given by the bonds C₅–C₆ and C₆–C₇. In the anionic model, the phenolic oxygen is deprotonated, and the oxygen–carbon bond, O₁–C₂, shortens by about 0.1 Å as compared to the neutral model, losing its single-bond character. As a consequence, the aromaticity of the phenolic ring is reduced, yielding a quinoid-like structure of the ring. The degree of bond alternation in the central carbon bridge is also decreased in the anionic chromophore, where the two central bond lengths differ by only 0.02 Å as compared to a value of about 0.08 Å in the neutral chromophore. Beyond the central bridge, the deprotonation of the phenolic oxygen does not have a large effect.

The neutral⁺ (E) chromophore represents a more complicated case due to the existence of two possible isomers. One isomer is depicted in Figure 1E and is characterized by the additional proton sitting on the NH₃⁺ group, hydrogen bonded to the imidazolone oxygen. In the other isomer, the proton has jumped on the imidazolone oxygen leaving behind a neutral NH₂ moiety. In the paper where this compound was first studied, only the existence of the first isomer was postulated, and a minimal geometry was also obtained using MP3 techniques.¹⁸ We find that, when optimizing the geometry of chromophore (E) with the BLYP functional, the structure with the proton on the NH₃⁺ group is not stable as the proton prefers to bind to the oxygen. The use of the B3LYP functional yields instead two minima with the proton bound to the oxygen being energetically slightly favored by 2 mHartree. The BLYP and B3LYP geometries with the proton bound on the imidazolone oxygen are practically identical. Finally, we note that the bond length pattern of the neutral⁺ chromophore has significant differences from both the anionic and the neutral models. Along the phenolic ring, the structure of the (E) model resembles the neutral form, but the carbon bridge is characterized by a much smaller bond-length alternation of only 0.03 Å, in line with the anionic value. Perhaps not surprisingly, the structure of the imidazolone ring of the (E) model differs from both the neutral and the anionic case due to the additional proton close to it.

5.2. TDDFT Excited States. We compute the low-lying singlet linear-response adiabatic TDDFT vertical excitations of the minimal (A, B), methyl-terminated (C, D), and neutral⁺ (E) models. The BLYP and B3LYP functionals are employed on the corresponding BLYP and B3LYP geometries using the cc-pVTZ basis, while the SAOP excitations are computed only on the BLYP geometries with the ET-pVQZ basis.

Table 1. Adiabatic TDDFT Vertical Excitation Energies (eV) and Oscillator Strengths (in Parentheses) for the Minimal (A) and Methyl-Terminated (B) Anionic Models^a

	BLYP	B3LYP	SAOP
Minimal Model (A)			
S ₀ → S ₁	2.30(0.00) H-1→L(0.70)	3.16(0.00) H-1→L(0.69)	2.59(0.00) H-1→L(1.00)
S ₀ → S ₂	2.97 (0.75) H→L(0.54)	3.16 (0.88) H→L(0.58)	2.99 (0.79) H→L(0.95)
-ε _{HOMO}	0.59	1.29	5.10
Methyl-Terminated Model (B)			
S ₀ → S ₁	2.29(0.00) H-1→L(0.70)	3.09 (0.92) H→L(0.58)	2.58(0.00) H-1→L(1.00)
S ₀ → S ₂	2.89 (0.77) H→L(0.54) H→L+2(0.13) H→L+3(0.11)	3.16(0.00) H-1→L(0.69)	2.93 (0.80) H→L(0.94)
-ε _{HOMO}	0.53	1.22	5.02

^aThe BLYP and B3LYP excitations are computed with a cc-pVTZ basis on the corresponding ground state geometries, while SAOP employs the BLYP geometries and the ET-pVQZ basis. The electronic transitions are listed if their weight (in parentheses) is greater than 0.1. The ionization threshold (-ε_{HOMO}) is also given. We use boldface characters to indicate the dominant excitation.

For all models, we report the TDDFT excitation energies with their oscillator strength and character. We also list minus the Kohn–Sham energy of the highest occupied molecular orbitals (HOMO) which gives the DFT ionization threshold. TDDFT may place a bound excitation in the continuum since the ionization threshold is underestimated in DFT due to the incorrect exponential decay of the Kohn–Sham potential for a pure functional.⁷⁴ The use of an asymptotically corrected exchange–correlation potential as SAOP cures this deficiency, raising the threshold to a significantly higher value, and allows us to verify that the excitation can in fact be trusted.⁷⁵

In Table 1, we summarize the TDDFT results for the anionic minimal (A) and methyl-terminated (B) models. As expected, the electronic properties of these two chromophores are rather similar. The addition of the methyl groups only lowers the BLYP, B3LYP, and SAOP dominant excitations by 0.08, 0.07, and 0.06 eV, respectively, while preserving the same HOMO → LUMO character of the excitation and the same oscillator strength. The excitations obtained with B3LYP are consistently higher by about 0.2 eV than the BLYP values as can be expected for a hybrid functional, while SAOP gives excitations which are slightly higher than BLYP. For the methyl-terminated (B) model, the TDDFT excitation lies in the range 2.9–3.1 eV and is therefore significantly different from the location at 2.59 eV of the absorption maximum of photodestruction spectroscopy experiments.¹⁶

For both models, the dominant TDDFT excitation is a π → π* transition and is not characterized by charge transfer as we can infer by inspecting the HOMO and LUMO orbitals in Figure 4. The HOMO orbital is π-bonding on both bonds of the central bridge and on the C₁₀–N₁₁ bond of the imidazolone ring, while the LUMO is antibonding on these three bonds. As these orbitals are rather similar for the (A) and (B) models and all functionals, we only plot them for SAOP approach and the methyl-terminated (B) model.

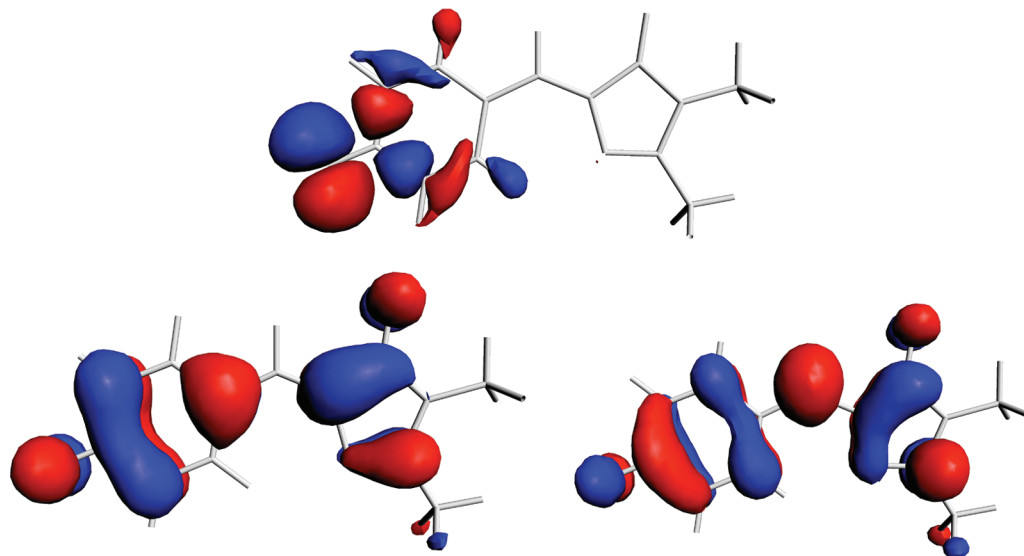


Figure 4. The DFT/SAOP orbitals for the anionic methyl-terminated (B) model. An isosurface of 0.03 is shown in red, and an isosurface of -0.03 is shown in blue. We only show the HOMO-1 (top), HOMO (bottom left), and LUMO (bottom right) orbitals, which are involved in the TDDFT/SAOP excitations.

Table 2. TDDFT Excitation Energies (eV) and Oscillator Strengths (in Parentheses) for the Minimal (C) and the Methyl-Terminated (D) Neutral Models^a

	BLYP	B3LYP	SAOP
Minimal Model (C)			
$S_0 \rightarrow S_1$	2.88(0.00)	3.54 (0.68)	3.13(0.00)
	H-1 \rightarrow L(0.70)	H \rightarrow L(0.61)	H-1 \rightarrow L(1.00)
$S_0 \rightarrow S_2$	3.22 (0.59)	3.56(0.00)	3.28 (0.62)
	H \rightarrow L(0.57)	H-1 \rightarrow L(0.69)	H \rightarrow L(0.95)
	H-5 \rightarrow L(0.13)		
$-\epsilon_{\text{HOMO}}$	4.98	5.85	9.50
Methyl-Terminated Model (D)			
$S_0 \rightarrow S_1$	2.99(0.00)	3.46 (0.66)	3.20 (0.61)
	H-1 \rightarrow L(0.70)	H \rightarrow L(0.61)	H \rightarrow L(0.94)
$S_0 \rightarrow S_2$	3.10 (0.52)	3.66(0.00)	3.21(0.00)
	H \rightarrow L(0.57)	H-1 \rightarrow L(0.69)	H-1 \rightarrow L(1.00)
	H-5 \rightarrow L(0.14)		
	H-2 \rightarrow L(0.13)		
$-\epsilon_{\text{HOMO}}$	4.75	5.62	9.27

^a See footnote a in Table 1 for further explanations.

We observe that the BLYP and B3LYP excitations of both anionic chromophores are significantly above the corresponding ionization threshold and lie in the continuum. The use of the SAOP potential cures the problem since the SAOP ionization threshold is well above the dominant excitation, which is now bound. Nevertheless, the TDDFT/SAOP excitation energies for the (A) and (B) models are very close to the BLYP values, and the picture remains practically unchanged with respect to the use of the BLYP functional.⁷⁵

The TDDFT results for the neutral minimal (C) and methyl-terminated (D) models are summarized in Table 2. As in the anionic case, the electronic properties of the two chromophores are rather similar. The dominant excitation has HOMO \rightarrow LUMO character and is only lowered by about 0.1 eV when the methyl groups are added. The B3LYP excitation energies are larger than the BLYP ones by more than 0.3 eV, while SAOP yields excitations which are at most 0.1 eV larger than the BLYP energies. The excitations of the neutral chromophores are larger than the excitations

of the corresponding anionic models by roughly 0.2, 0.4, and 0.3 eV for the BLYP, B3LYP, and SAOP functionals, respectively.

Differently from the case of the anionic models, the underestimation of the DFT ionization threshold is not a concern for the neutral chromophores as this threshold is above the dominant excitation for both the BLYP and B3LYP functionals. The character of the excitations is $\pi \rightarrow \pi^*$ also in the neutral models as can be seen from the orbitals in Figure 5. We only plot the orbitals which are involved in the SAOP excitations for the neutral methyl-terminated (D) model since these orbitals are rather similar for the other model and the other functionals. The HOMO orbital has π -antibonding character on the first carbon bond of the central bridge, while it is π -bonding on the second bond. The situation is reversed in the LUMO where the sequence along the bridge is instead π -bonding/antibonding.

In Table 3, we report the TDDFT excitations for the neutral⁺ (E) model. The B3LYP excitations for the two structural B3LYP minima with the proton on the imidazolone oxygen and with the proton on the NH_3^+ group only differ by 0.03 eV, so the two structures are energetically very close both in the ground and in the excited state. For the structure with the proton on the oxygen, BLYP yields an excitation energy lower by about 0.2 eV than the B3LYP value, while SAOP lies as always between BLYP and B3LYP. We note that, for this model, the DFT underestimation of the ionization threshold is not an issue. Photodestruction spectroscopy experiments place the absorption maximum at 2.99 eV,¹⁸ reasonably close to the vertical BLYP excitation but 0.2 and 0.3 eV lower than the SAOP and B3LYP values, respectively.

5.3. CASPT2 Calculations. The CASPT2 vertical excitation energies are computed on the BLYP geometries of all models in Figure 1. Unless otherwise stated, the standard IPEA zero-order Hamiltonian⁶⁵ is used, and no additional constant level shift⁶⁶ is applied.

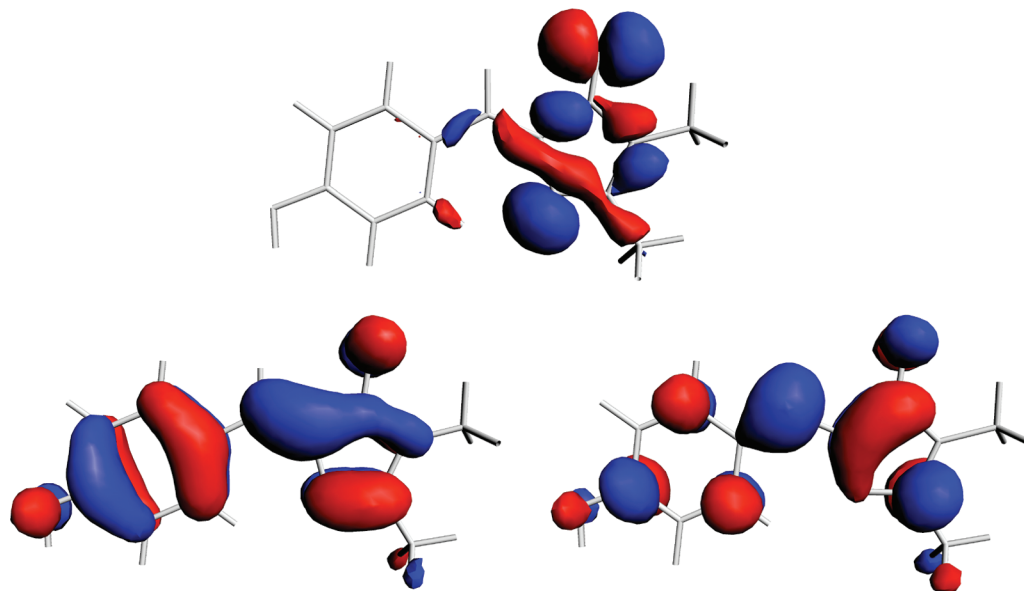


Figure 5. The DFT/SAOP orbitals for the neutral methyl-terminated (D) model. An isosurface of 0.03 is shown in red, and an isosurface of -0.03 is shown in blue. We only show the HOMO-1 (top), HOMO (bottom left), and LUMO (bottom right) orbitals, which are involved in the TDDFT/SAOP excitations.

Table 3. Adiabatic TDDFT Vertical Excitation Energies (eV) and Oscillator Strengths (in Parentheses) for the Neutral⁺ (E) Model^a

	BLYP	B3LYP	B3LYP	SAOP
$S_0 \rightarrow S_1$	H ⁺ on O 3.07 (0.76)	H ⁺ on O 3.30 (0.92)	H ⁺ on N 3.33 (0.91)	H ⁺ on O 3.21 (0.69)
	H \rightarrow L(0.55)	H \rightarrow L(0.59)	H \rightarrow L(0.60)	H \rightarrow L(0.83)
	H-1 \rightarrow L(0.15)			H-1 \rightarrow L(0.15)
$-\epsilon_{\text{HOMO}}$	8.29	9.17	8.74	12.79

^a For the B3LYP functional, we list the results for the two geometries with the proton on the imidazolone oxygen (H⁺ on O) and with the proton on the NH₃⁺ group (H⁺ on N). See footnote a in Table 1 and text for further explanations.

Table 4. CASPT2 Excitation Energies (eV) of the Anionic Minimal Model (A) Obtained Using Different Basis Sets and CAS(n, m) Expansions with n Electrons Distributed over m Orbitals of a'' Symmetry (π Orbitals)^a

CAS(n, m)	6-31G*	cc-pVDZ	cc-pVTZ
2,2	2.84	2.81	2.70
4,4	3.06	3.06	2.97
6,6	3.00	3.01	2.93
8,8	2.99	3.01	2.91
10,10	2.98	3.00	2.92
12,12	2.94	2.96	2.87
14,14	2.95	2.96	2.88
16,15	2.92	-	-

^a The number of doubly occupied a'' orbitals in the Hartree-Fock reference configuration is 8 for a total of 16 electrons. The DFT/BLYP ground state geometry is used.

The CASPT2 energies of the minimal anionic (A) model are listed in Table 4 for increasing size of the CAS expansion and different basis sets. The ground and the dominant excited state have A' character, and the CAS(n, m) expansion to describe both states in a state-average CASSCF calculation is obtained by distributing n electrons over m orbitals of a'' symmetry, that is, of π bonding/antibonding character. The system has 8 doubly occupied a'' orbitals in the Hartree-Fock

reference configuration, so 16 is the maximum number of π electrons which can be excited. We employ the cc-pVDZ and cc-pVTZ basis sets as well as the small 6-31G* basis set also used in the CASPT2/CASSCF(12,11) calculations on the same model in ref 37.

The dependence of the CASPT2 excitation energy on the size of the CAS is not very strong, and the result is already converged to better than 0.1 eV when a small CAS(6,6) expansion is employed. The character of the ground state is not strongly multiconfigurational, and the excited state is predominantly a HOMO \rightarrow LUMO excitation. The oscillator strength computed using the state-average CASSCF wave functions is always greater than one. The 6-31G* and cc-pVDZ basis sets yield comparable excitations, while employing the cc-pVTZ basis systematically lowers the excitation by roughly 0.1 eV.

Our CASPT2 calculations therefore place the vertical excitation energy of the minimal anionic (A) model at about 2.9 eV. This result is at variance with previous CASPT2 calculations³⁷ which find an excitation energy of 2.67 eV employing the 6-31G* basis and a CASSCF(12,11) expansion. Our tests indicate that the basis and the size of the CAS do not influence the result to such an extent and that the discrepancy originates from the use of two different zero-order Hamiltonians. We employ the standard IPEA Hamiltonian implemented in MOLCAS 7.2, while the authors of ref 37 use a previous version of the code where the IPEA Hamiltonian was not available. If we perform a CASPT2 calculation with the CAS(14,14) expansion and set the IPEA shift to zero to reproduce similar conditions to ref 37, we find indeed an excitation of 2.63 eV.

For the methyl-terminated anionic (B) model, we find that CASPT2 predicts a very similar behavior as in the case of the minimal (A) model. The character and CASSCF oscillator strength of the electronic states are comparable, and, as a function of CAS size, the CASPT2 excitation energies are

Table 5. MS-CASPT2 Excitation Energies (eV) and CASSCF Oscillator Strengths (in Parentheses) of the Neutral Minimal (C) Model Obtained Using the cc-pVDZ Basis Set and CAS(n,m) Expansions with n Electrons Distributed over m Orbitals of a'' Symmetry (π Orbitals)^a

CAS(n,m)	energy (osc. strength)
10,10	4.13 (0.96)
12,12	4.03 (1.00)
14,13	3.77 (0.82)
14,14	3.78 (0.78)
16,15	3.72 (0.85)

^aWe use 4 states in the state-average CASSCF and MS-CASPT2 calculations, and the dominant state is always the 4th within CASSCF and the 2nd state within MS-CASPT2. We apply an additional level shift of 0.1 hartree in all CASPT2 calculations except in the CAS(16,15) calculation. The DFT/BLYP ground state geometry is used.

only slightly lower than the values of the minimal (A) model of Table 4. The CASPT2 excitation energy of the methyl-terminated anionic (B) model computed with a CAS(14,14) and a cc-pVDZ basis is equal to 2.92 eV. Photodestruction spectroscopy experiments locate the absorption maximum for this model at 2.59 eV.¹⁶

In Table 5, we show the CASPT2 excitation energies for the neutral minimal (C) model. The nature of the electronic states is different and more complex than in the anionic case since the dominant excited state is now strongly multiconfigurational and the CASPT2 result is very sensitive to the size of the CAS expansion. Moreover, the dominant CASSCF excited state is not always the lowest, but, as the size of the CAS increases, it becomes one of three excited states with relatively close CASSCF energies in a range of only 30 mHartree. Therefore, we need to perform multistate (MS) CASPT2 calculations over more than two states and to introduce an additional level shift to ensure convergence of the CASPT2 result. In the MS-CASPT2 calculations, the dominant excited state is the second one and is separated by more than 30 mHartree from the subsequent higher states, so the near degeneracy observed in CASSCF is lifted. In Table 5, we do not report the results for the smaller CAS expansions as the variations in the CASPT2 energy of the dominant state are as large as 0.3 eV as a function of the number of states included and the magnitude of the added level shift. Only with an expansion as large as a CAS(10,10) over 4 states and with a level shift of 0.1 hartree, we obtain a stable result in the sense that the dominant CASSCF state is the fourth one, an increase of the shift does not significantly affect the result, and employing 5 states with a shift of 0.2 hartree changes the excitation by only 0.02 eV. We note that this problematic convergence was ignored in a previous CASPT2 study of the neutral form using only a CAS(2,2) expansion over 2 states.⁴¹ From Table 5, we observe that the convergence of the CASPT2 energy is not as smooth as in the anionic case, and it is necessary to correlate as many as 14 electrons to yield an apparently converged MS-CASPT2 result. As for the anionic case, we test the effect of setting the IPEA shift to zero and find that the MS-CASPT2 excitation with a CAS(14,14) expansion is lowered by as much as 0.45 eV when compared to the value given in Table 5.

Table 6. EOM-CCSD Vertical Excitation Energies in eV of the Anionic Minimal (A) and Methyl-Terminated (B), Neutral Minimal (C) and Methyl-Terminated (D), and Neutral⁺ (E) Models^a

model	cc-pVDZ	cc-pVTZ'
minimal anionic (A)	3.08	3.03
minimal neutral (C)	4.11	4.01
methyl-term. anionic (B)	3.04	-
methyl-term. neutral (D)	4.00	-
neutral ⁺ (E)	3.31	-

^aWe employ the cc-pVDZ and cc-pVTZ' basis sets. The DFT/BLYP ground state geometries are used.

We also employ a CAS(14,14) expansion over 4 states to compute the excitation energy of the methyl-terminated neutral (D) model with the MS-CASPT2 approach and a level shift of 0.1 hartree. The resulting excitation of 3.58 eV is 0.2 eV lower than the CAS(14,14) value obtained for the minimal (C) model. Finally, we note that the excitation of the methyl-terminated neutral (D) model is significantly higher by as much as 0.7 eV than the excitation of the anionic (B) model.

Finally, we compute the CASPT2 excitation of the neutral⁺ (E) model, using the BLYP geometry where the additional proton is bound to the oxygen of the imidazolone ring. It is not a priori obvious whether the behavior of the electronic states of the neutral⁺ model are closer to the neutral or the anionic case since its geometrical features are distinct from either of them as discussed in Section 5.1. We find that the dependence of the CASPT2 excitation on the size of the CAS expansion is very smooth and similar to what we observe for the anionic model. The dominant excitation has strong HOMO \rightarrow LUMO character, and a CAS(6,6) yields already a converged excitation to better than 0.1 eV. The CASPT2 excitation computed with a CAS(14,14) expansion and a cc-pVDZ basis is equal to 3.21 eV. Also in this case, setting the IPEA shift to zero reduces the excitation energy to 2.87 eV. We recall that photodestruction experiments locate the absorption maximum of the neutral⁺ chromophore at 2.99 eV.¹⁸

5.4. CC Calculations. We compute the EOM-CCSD vertical excitation energies of all models in Figure 1. We employ the BLYP ground state geometries and the cc-pVDZ basis. For the (A) and (C) models, we also use the cc-pVTZ' basis set.

The EOM-CCSD results are shown in Table 6. For all systems, the state with the largest oscillator strength is the second one and displays HOMO \rightarrow LUMO character. We find that the use of a larger basis on the heavy atoms lowers the excitations of the minimal anionic (A) and neutral (C) models by only 0.05 and 0.1 eV, respectively. As expected, the excitation of the methyl-terminated anionic (B) model is very close to the value of the minimal anionic (A) model. The effect of deprotonation is instead very large as the difference between the excitation energy of the neutral (C) and anionic (A) chromophores is about 1.0 eV. The same large difference is observed between the excitations of the neutral and anionic methyl-terminated chromophores. Our EOM-CCSD results for the minimal models are in agreement with previous CC calculations.⁴¹ Finally, the excitation

Table 7. Variational (VMC) and Diffusion Monte Carlo (DMC) Vertical Excitations in eV of the Minimal Anionic (A) Model, Computed with the cc-pVDZ Basis Set^a

CAS(<i>n,n</i>)	Thr _{CSF}	N _{CSF}	VMC	DMC
2,2	0.00	3	3.41(4)	3.21(4)
4,4	0.00	20	3.39(4)	3.18(4)
6,6	0.10	7	3.18(4)	3.11(4)
8,8	0.10	9	3.14(4)	3.02(4)
8,8	0.05	25	3.18(4)	3.07(4)
8,8	0.02	75	3.21(4)	3.04(4)
10,10	0.05	27	3.20(4)	3.06(4)

^a The statistical error on the last digit is indicated in brackets. We employ CAS(*n,n*) expansions obtained by distributing *n* electrons over *n* orbitals of *a''* symmetry and expressed on the state-averaged CASSCF natural orbitals. We list the threshold applied to the CSF coefficients and the number of CSFs kept in the determinantal component of the QMC wave function. The CI and Jastrow parameters are optimized in a state-average minimization within VMC. The DFT/BLYP ground state geometry is used.

energy of the neutral⁺ (E) is quite different than the value of the neutral model.

5.5. QMC Calculations. We compute the QMC vertical excitation energies of the minimal anionic (A), methyl-terminated (B), and neutral⁺ (E) models. We employ the BLYP ground state geometries and the cc-pVDZ basis.

In Table 7, we present the VMC and DMC excitation energies of the minimal anionic (A) model, computed with different determinantal components in the trial wave function where only the CI coefficients and Jastrow parameters are optimized in a state-averaged minimization within VMC. As in the CASPT2 calculations, we compute the QMC excitations using CAS(*n,n*) expansions of increasing size obtained by distributing *n* electrons over *n* orbitals of *a''* symmetry. We do not consider expansions larger than CAS(10,10) since we have already demonstrated that the states of the anionic chromophore do not have a strong multiconfigurational character.

We notice that, starting with the CAS(6,6) expansion, the VMC excitation energy is reduced by about 0.2 eV with respect to the smaller CAS wave functions. The DMC excitation is also lowered by roughly the same amount when a CAS(8,8) is employed. In the CAS(8,8) calculations, we use three different thresholds to truncate the expansion and find equivalent VMC excitations in the range of 3.14–3.21 eV and in agreement with the CAS(6,6) value. The three corresponding DMC excitation energies are also compatible within one statistical error and fall in the range of 3.02–3.07 eV. Using a larger CAS(10,10) expansion truncated with a threshold of 0.05 yields a DMC excitation energy consistent with the CAS(8,8) results.

For the minimal anionic (A) model, we also investigate the effect of optimizing the orbitals together with the CI and Jastrow parameters and perform a full state-averaged optimization within VMC for the CAS(2,2) and CAS(4,4) wave functions as well as for the CAS(6,6) and CAS(8,8) wave functions truncated with a threshold of 0.1. Even though the gain in the VMC energy for each state can be as large as 30 mHartree, we find that, in all cases, the VMC and DMC excitation energies obtained with fully optimized QMC wave functions are compatible within one standard deviation with

Table 8. Vertical Excitations Energies (eV) of the Minimal Anionic (A) and Neutral (C), Methyl-Terminated Anionic (B) and Neutral (D), and Neutral⁺ (E) Models^a

model	DMC	SAOP	CASPT2	CCSD	PDS expt.	sol. expt.
anionic (A)	3.06(4)	2.99	2.96	3.08	-	-
anionic (B)	3.04(4)	2.93	2.92	3.04	2.59 ^b	2.84 ^d
neutral (C)	-	3.28	3.72	4.11	-	-
neutral (D)	-	3.20	3.58	4.00	-	3.51 ^d
neutral ⁺ (E)	3.36(5)	3.21	3.21	3.31	2.99 ^c	-

^a We list the results computed with diffusion Monte Carlo (DMC), TDDFT/SAOP, CASPT2/cc-pVDZ, and EOM-CCSD/cc-pVDZ. The ground state DFT/BLYP geometries are used. We report the photodestruction spectroscopy (PDS) experimental absorption maxima from refs 16^b and 18^c as well as the results of the Kamlet-Taft multivariate fit of the absorption maxima in various solvents^{22d}.

the corresponding values obtained by optimizing only the CI and Jastrow parameters.

For the methyl-terminated anionic (B) model, we compute the VMC and DMC excitation energies using wave functions obtained from a CAS(8,8) expansion truncated with a threshold of 0.05. We do not employ larger CAS expansions as we know that the CASSCF and CASPT2 calculations yield a well converged excitation energy already with a CAS(8,8) as discussed in Section 5.3. Moreover, the QMC calculations for the minimal (A) model indicate that the DMC excitations are well converged with this choice of CAS and threshold. The VMC and DMC excitation energies for the methyl-terminated anionic (B) model are equal to 3.21(4) and 3.04(4) eV, respectively.

Finally, we compute the QMC excitation energies of the neutral⁺ (E) model using wave functions obtained from a CAS(8,8) expansion truncated with a threshold of 0.05 on the CSF coefficients. We find that the VMC and DMC excitation energies are equal to 3.43(5) and 3.36(5) eV, respectively.

6. Discussion and Conclusions

We will begin with the analysis of the relative performance of the various theoretical approaches and then compare with available experimental data. In particular, we will focus on the results from photodestruction spectroscopy experiments¹⁶ which predict the bathochromic shift to be very small when the anionic chromophore is embedded in the protein as well as on the data in solution²² where instead the absorption maximum of the neutral form is found to be significantly blue-shifted with respect to the protein.

To simplify the discussion, we summarize the key results in Table 8, where we report the vertical excitations of the minimal anionic (A) and neutral (C), the methyl-terminated anionic (B) and neutral (D), and the neutral⁺ models. We recall that the minimal (A) and (C) models are sufficient to access the impact of deprotonation on the excitation energy, while, for the methyl-terminated (B) and neutral⁺ (E) chromophores, photodestruction spectroscopy experiments are available.^{16,18} For the methyl-terminated (B) and (D) models, we can also compare to absorption experiments in solution.²² We list the TDDFT results obtained using the

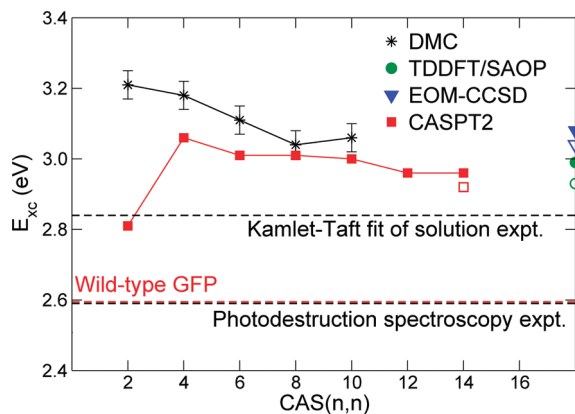


Figure 6. DMC and CASPT2 excitation energies (eV) of the anionic minimal (A) model obtained using increasing CAS(n , n) expansions. The EOM-CCSD and TDDFT/SAOP excitations are also reported on the right axis where the filled and open symbols refer to the anionic minimal (A) and the anionic methyl-terminated (B) model, respectively. The CASPT2 energy computed with the CAS(14,14) is also plotted for the (B) model using an open square symbol. The DFT/BLYP ground state geometry is used. For the methyl-terminated anionic (B) model, we indicate the location of the absorption maximum in photodestruction spectroscopy experiments¹⁶ and the result of the Kamlet-Taft multivariate fit of the absorption maxima in various solvents.²² The value of the absorption maximum in the wild-type GFP²⁹ is also plotted.

SAOP approach since SAOP yields excitations which are between the BLYP and B3LYP results and cures the problem of the underestimation of the ionization threshold in the anionic models. For the CASPT2 and EOM-CC excitations, we only list the values obtained with the cc-pVDZ basis since results with this basis are consistently available for all models. We recall that the use of a larger basis set in these calculations is found to lower the excitation energies by roughly 0.05–0.1 eV. For the CASPT2 and DMC excitations, we report the values obtained with the largest CAS expansions.

We first observe that all theoretical approaches agree in predicting an excitation energy of the minimal anionic (A) model in the range of 2.96–3.08 eV. It is evident from Figure 6 that, when the quality of the wave function is improved, CASPT2 and DMC converge to a similar excitation, which is also in close agreement with the TDDFT/SAOP and EOM-CC values. We already noted that our CASPT2 excitation energy is significantly higher than the CASPT2 value of 2.67 eV reported in ref 37, and we attributed this discrepancy to the use of a different zero-order Hamiltonian. The IPEA Hamiltonian used in this work yields on average more accurate excitations, and its good performance appears to be here corroborated by the good agreement found between the CASPT2 and the DMC and EOM-CC excitations.

For all theoretical methods, the excitation energies of the methyl-terminated anionic (B) model are only slightly smaller than the corresponding values for the minimal anionic (A) model by about 0.05 eV. This is not surprising since the two chromophores only differ in the termination. We can therefore conclude that theory predicts an excitation of the methyl-terminated (B) model in the range of 2.92–3.04 eV.

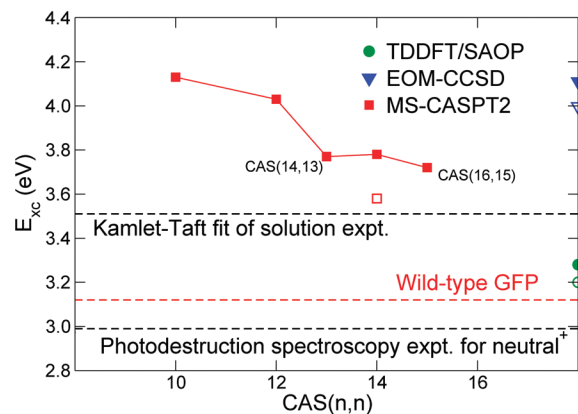


Figure 7. MS-CASPT2 excitation energies (eV) of the minimal neutral (C) model obtained using increasing CAS(n , n) expansions and 4 states. The EOM-CCSD and TDDFT/SAOP excitations are also reported on the right axis where the open symbols refer to the methyl-terminated (D) model. The MS-CASPT2 energy computed with the CAS(14,14) is also plotted for the (D) model using an open square symbol. For the methyl-terminated neutral (D) model, we indicate the result of the Kamlet-Taft multivariate fit of the absorption maxima in various solvents.²² The value of the room-temperature absorption maximum in the wild-type GFP²⁹ is also plotted together with the location of the absorption peak of the neutral⁺ (E) model measured in photodestruction spectroscopy experiments.¹⁸

This estimate is rather robust since the effect of using a larger basis only affects the CASPT2 and EOM-CC excitations of the smaller minimal (A) model by less than 0.1 eV, and we may expect a similar behavior in the case of methyl terminations.

The minimal neutral (C) chromophore appears to be more difficult and controversial for theoretical methods than the anionic case. TDDFT/SAOP yields an excitation energy significantly smaller than the CASPT2 and EOM-CC values, which do also not agree with each other. TDDFT/SAOP places the vertical excitation of the neutral chromophore only 0.29 eV higher than the excitation of the anionic counterpart, while CASPT2 and EOM-CC predict a significant blue shift of 0.76 and 1.02 eV, respectively. The more complex nature of the excitation in the neutral chromophore is evident from Figure 7, where the convergence of the CASPT2 energy with the dimension of active space is clearly nontrivial: We employ the MS-CASPT2 approach over 4 states and do not even show the results for small CAS expansions given their dramatic dependence on the details of the calculations.

For the neutral⁺ (E) model, all theoretical methods agree within 0.1 eV and predict an excitation energy in the range of 3.21–3.36 eV and, therefore, a similar shift of 0.27–0.32 eV between the excitations of the (E) and the anionic methyl-terminated (B) model. As it was the case for the anionic chromophore, the neutral⁺ model does not pose any particular difficulty in the calculations. The states are not strongly multiconfigurational, and the CASPT2 and DMC excitations are well converged already when a small CAS expansion is employed. This must be contrasted to the situation encountered in the study of the neutral model. Therefore, even though the neutral⁺ model was originally synthesized to

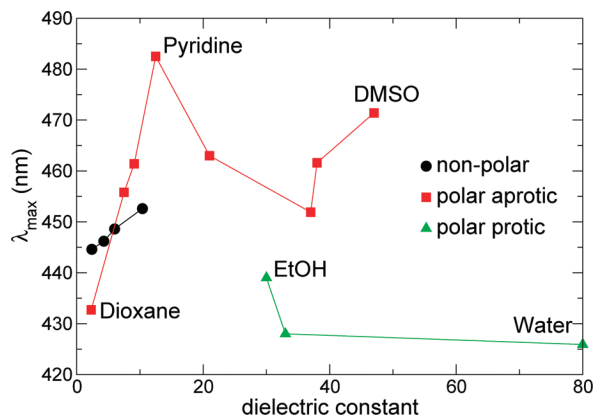


Figure 8. Experimental absorption maxima (nm) of the methyl-terminated anionic (B) model in solution from ref 22. The location of the peak is plotted as a function of the dielectric constant of the solvent. The solvents are classified as nonpolar, polar aprotic, and polar protic.

mimic the behavior of the neutral chromophore in photodestruction spectroscopy experiments,¹⁸ its electronic states are very perturbed by the presence of the additional proton, and, consequently, this chromophore shares in fact very little with the neutral one.

To understand the link between our theoretical calculations and experiments, we report the relevant experimental data in Figures 6 and 7 as well as in Table 8. For the anionic chromophore, all theoretical approaches yield a vertical excitation energy in the gas phase significantly higher than the room-temperature absorption peak at 2.59 eV of the anionic form in the wild-type GFP protein.²⁹ Therefore, the predicted bathochromic shift is large, of the order of 0.3–0.4 eV, and the protein pocket must significantly modulate the excitation with respect to vacuum. This is in contrast to photodestruction spectroscopy experiments in the gas phase which predict an excitation very close to the value in the protein.¹⁶ The disagreement between our theoretical results and photodestruction spectroscopy experiments is puzzling.

In an attempt to clarify this discrepancy, we also analyze recent extensive experimental data in solution²² and plot the location of the absorption maximum of the anionic chromophore as a function of the dielectric constant of the solvent in 8. We note that the location of the absorption peak strongly depends on the solvent and spans a range as large as 60 nm (0.34 eV). While some solvents such as DMSO give very red-shifted excitations close to the value of 478 nm (2.59 eV) in the protein,¹⁶ we can see at a glance that nonpolar solvents yield significantly blue-shifted values, and the extrapolation to the vacuum dielectric constant of one falls in the range 440 ± 5 nm (2.82 ± 0.03 eV). A more quantitative analysis is carried out in ref 22, where the multivariate Kamlet-Taft fit⁷⁶ of the absorption maxima in terms of the acidic, basic, and polar solvating parameters extrapolates to zero, that is, to the conditions of vacuum, at 2.84 eV. This value is only slightly lower than the vertical excitations computed with the various theoretical approaches, and this small red shift of 0.1–0.2 eV can easily be explained with a combination of basis set errors and vibronic effects which are very strong in this system. To estimate the energetic shift in the excited state due to vibronic effects,

we perform a CASPT2 geometry optimization in the excited state, preserving C_s symmetry since the initial excited-state relaxation is believed to occur mainly in plane.^{20,23} We find that the excitation is lowered by about 0.1 eV, a value compatible with a previous CASPT2/CASSCF study.³⁷

For the neutral chromophore, the multivariate Kamlet-Taft fit of the absorption maxima in solution extrapolates to the conditions of vacuum, at 3.51 eV.²² This value is significantly blue-shifted with respect to the absorption at 3.12 eV of the neutral form in the wild-type GFP protein, and no solvent yields an absorption sufficiently red-shifted to cover the value in the protein.²⁹ Our CASPT2 calculations support the picture emerging from the data in solution and yield an excitation of about 3.6 eV, slightly blue-shifted with respect to the extrapolation from the solution data but still perfectly compatible if basis set and vibronic effects are taken into account as in the case of the anionic chromophore. For the neutral case, there are no gas-phase experiments available, and we note that the absorption maximum observed in photodestruction spectroscopy experiments for the neutral⁺ model¹⁸ is rather far from the theoretical estimate obtained for the neutral chromophore and even slightly below the absorption value in the protein. We already pointed out that, theoretically, the neutral⁺ model has a significantly lower excitation and little in common with the neutral chromophore in vacuum. Interestingly, we may expect this model to be closer to the situation in the protein where the Arg-96 positive counterion is in close proximity to the imidazolone oxygen. Therefore, the significant red-shift of the neutral⁺ absorption in the direction of the protein value can be explained with the fact that neutral⁺ does not mimic the neutral form in vacuum as originally suggested but possibly mimics the neutral chromophore in the protein environment.

In summary, we employ state-of-the-art theoretical approaches to investigate the absorption properties of several GFP chromophore models in the gas phase. Our theoretical findings nicely support the picture emerging from absorption experiments in solution and raise concerns on the interpretation of gas-phase photodestruction spectroscopy experiments especially for the anionic case. Our theoretical vertical excitations in the gas phase clearly indicate that the bathochromic shifts are quite significant and that the conditions of the chromophore in the protein are not close to those in vacuo. We finally note that our CASPT2 calculations differ from previous studies in the use of a more accurate zero-order Hamiltonian and that TDDFT differs significantly from highly correlated approaches in treating the neutral form of the chromophore.

Acknowledgment. We thank Mark Casida and Claudio Amovilli for many useful discussions. M.Z. is supported by the Stichting voor Fundamenteel Onderzoek der Materie (FOM). We acknowledge support from the Stichting Nationale Computerfaciliteiten (NCF-NWO) for the use of the SARA supercomputer facilities.

Supporting Information Available: DFT geometries of the model chromophores and CASPT2 geometry of the anionic minimal (A) model. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) *Time-dependent Density Functional Theory, Lecture Notes of Physics*; Marques, M. A. L., Ullrich, C. A., Nogueira, F., Rubio, A., Burke, K., Gross, E. K. U., Eds.; Springer: Berlin, 2006.
- (2) Tozer, D. J.; Amos, R. D.; Handy, N. C.; Roos, B. O.; Serrano-Andrés, L. *Mol. Phys.* **1999**, *97*, 859–868.
- (3) Casida, M. E.; Gutierrez, F.; Guan, J.; Gadea, F.-X.; Salahub, D. R.; Daudey, J.-P. *J. Chem. Phys.* **2000**, *113*, 7062–7071.
- (4) Dreuw, A.; Weisman, J. L.; Head-Gordon, M. *J. Chem. Phys.* **2003**, *119*, 2943–2946.
- (5) Casida, M. E. In *Recent Advances in Density Functional Methods, Part I*; Chong D. P., Ed.; World Scientific: Singapore, 1995; pp 155–188.
- (6) Maitra, N. T.; Zhang, F.; Cave, F. J.; Burke, K. *J. Chem. Phys.* **2004**, *120*, 5932–5937.
- (7) Cave, R. J.; Zhang, F.; Maitra, N. T.; Burke, K. *Chem. Phys. Lett.* **2004**, *389*, 39–42.
- (8) Casida, M. E. *J. Chem. Phys.* **2005**, *122*, 054111–1–8.
- (9) Roos, B. O.; Andersson, K.; Fülscher, M. P.; Malmqvist, P.-A.; Serrano-Andrés, L. In *Advances in Chemical Physics*; Prigogine, I., Rice, S. A., Eds.; Wiley & Sons, Inc.: New York, 1996; Vol. 93, pp 219–331.
- (10) Strambi, A.; Coto, P. B.; Frutos, L. M.; Ferré, N.; Olivucci, M. *J. Am. Chem. Soc.* **2008**, *130*, 3382–3388.
- (11) Sinicropi, A.; Martin, E.; Ryazantsev, M.; Helbing, J.; Briand, J.; Sharma, D.; Léonard, J.; Haake, S.; Cannizzo, A.; Chergui, M.; Zanirato, V.; Fusi, S.; Santoro, F.; Basosi, R.; Ferré, N.; Olivucci, M. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 17642–17647.
- (12) Shimomura, O.; Johnson, F. H.; Saiga, Y. *J. Cell. Comp. Physiol.* **1962**, *59*, 223–239.
- (13) Tsien, R. Y. *Annu. Rev. Biochem.* **1998**, *67*, 509–544.
- (14) Zimmer, M. *Chem. Rev.* **2002**, *102*, 759–781.
- (15) Tozzini, V.; Pellegrini, V.; Beltram, F. In *Handbook of organic photochemistry and photobiology*; Horspool, W. M., Lenci, F., Eds.; CRC Press: Washington, DC, 2004; Chapter 139, pp 1–23.
- (16) Nielsen, S. B.; Lapierre, A.; Andersen, J. U.; Pedersen, U. V.; Tomita, S.; Andersen, L. H. *Phys. Rev. Lett.* **2001**, *87*, 228102–1–4.
- (17) Andersen, L. H.; Bluhme, H.; Boyé, S.; Jorgensen, T. J. D.; Krogh, H.; Nielsen, I. B.; Nielsen, S. B.; Svendsen, A. *Phys. Chem. Chem. Phys.* **2004**, *6*, 2617–2627.
- (18) Lammich, L.; Pertersen, M. A.; Nielsen, M. B.; Andersen, L. H. *Biophys. J.* **2007**, *92*, 201–207.
- (19) Niwa, H.; Inouye, S.; Hirano, T.; Matsuno, T.; Kojima, S.; Kubota, M.; Ohashi, M.; Tsujii, F. I. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 13617–13622.
- (20) Mandal, D.; Tahara, T.; Webber, N. M.; Meech, S. R. *Chem. Phys. Lett.* **2002**, *358*, 495–501.
- (21) Vengris, M.; van Stokkum, I. H. M.; He, X.; Bell, A. F.; Tonge, P. J.; van Grondelle, R.; Larsen, D. S. *J. Phys. Chem. A* **2004**, *108*, 4587–4598.
- (22) Dong, J.; Solntsev, K. M.; Tolbert, L. M. *J. Am. Chem. Soc.* **2006**, *128*, 12038–12039.
- (23) Stavrov, S. S.; Solntsev, K. M.; Tolbert, L. M.; Huppert, D. *J. Am. Chem. Soc.* **2006**, *128*, 1540–1546.
- (24) Webber, N. M.; Meech, S. R. *Photochem. Photobiol. Sci.* **2007**, *6*, 976–981.
- (25) Chattoray, M.; King, B. A.; Bublitz, G. U.; Boxer, S. G. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 8362–8367.
- (26) Ormö, M.; Cubitt, A. B.; Kallio, K.; Gross, L. A.; Tsien, R. Y.; Remington, S. J. *Science* **1996**, *273*, 1392–1395.
- (27) Brejc, K.; Sixma, T. K.; Kitts, P. A.; Kain, S. R.; Tsien, R. Y.; Ormö, M.; Remington, S. J. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 2306–2311.
- (28) Bublitz, G. U.; King, B. A.; Boxer, S. G. *J. Am. Chem. Soc.* **1998**, *120*, 9370–9371.
- (29) Creemers, T. M. H.; Lock, A. J.; Subramaniam, V.; Jovin, T. M.; Völker, S. *Nat. Struct. Biol.* **1999**, *6*, 557–560.
- (30) Bell, A. F.; He, X.; Wachter, R. M.; Tonge, P. J. *Biochemistry* **2000**, *39*, 4423–4431.
- (31) Voityuk, A. A.; Michel-Beyerle, M. E.; Rösch, N. *Chem. Phys. Lett.* **1997**, *272*, 162–167. *Chem. Phys.* **1998**, *231*, 13–25. *Chem. Phys. Lett.* **1998**, *296*, 269–276.
- (32) Weber, W.; Helms, V.; McCammon, J. A.; Langhoff, P. W. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 6177–6182.
- (33) Laino, T.; Nifosi, R.; Tozzini, V. *Chem. Phys.* **2004**, *298*, 17–28.
- (34) Helms, V.; Winstead, C.; Langhoff, P. W. *J. Mol. Struct.* **2000**, *506*, 179–189.
- (35) El Yazal, J.; Prendergast, F. G.; Shaw, D. E.; Pang, Y.-P. *J. Am. Chem. Soc.* **2000**, *122*, 11411–11415.
- (36) Marques, M. A. L.; López, X.; Varsano, D.; Castro, A.; Rubio, A. *Phys. Rev. Lett.* **2003**, *90*, 258101–1–4.
- (37) Martin, M. E.; Negri, F.; Olivucci, M. *J. Am. Chem. Soc.* **2004**, *126*, 5452–5464.
- (38) Sinicropi, A.; Andruniow, T.; Ferré, N.; Basosi, R.; Olivucci, M. *J. Am. Chem. Soc.* **2005**, *127*, 11534–11535.
- (39) Vendrell, O.; Gelabert, R.; Moreno, M.; Lluch, J. M. *J. Am. Chem. Soc.* **2006**, *128*, 3564–3574.
- (40) Das, A. K.; Hasegawa, J.-Y.; Miyahara, T.; Ehara, M.; Nakatsuij, H. *J. Comput. Chem.* **2003**, *24*, 1421–1431.
- (41) Toniolo, A.; Olsen, S.; Manohar, L.; Martínez, T. J. *Faraday Discuss.* **2004**, *127*, 149–163.
- (42) Xie, D.; Zeng, X. *J. Comput. Chem.* **2005**, *26*, 1487–1496.
- (43) Nemukhin, A. V.; Topol, I. A.; Burt, S. K. *J. Chem. Theory Comput.* **2006**, *2*, 292–299.
- (44) Wan, S.; Liu, S.; Zha, G.; Chen, M.; Han, K.; Sun, M. *Biophys. Chem.* **2007**, *129*, 218–223.
- (45) Hasegawa, J.-Y.; Fujimoto, K.; Swerts, B.; Miyahara, T.; Nakatsuij, H. *J. Comput. Chem.* **2007**, *28*, 2443–2452.
- (46) Bravaya, K. B.; Bochenkova, A. V.; Granovskii, A. A.; Nemukhin, A. V. *Russ. J. Phys. Chem. B* **2008**, *2*, 671–675.
- (47) Monkhorst, H. J. *Intern. J. Quantum Chem. Symp.* **1977**, *11*, 421. Dalgaard, E.; Monkhost, H. J. *Phys. Rev. A* **1983**, *28*, 1217–1222.
- (48) Foulkes, W. M. C.; Mitas, L.; Needs, R. J.; Rajagopal, G. *Rev. Mod. Phys.* **2001**, *73*, 33–83.
- (49) *Molecular Electronic-Structure Theory*; Helgaker, T., Jorgensen, P., Olsen, J., Eds.; John Wiley and Sons, Inc.: Chichester, West Sussex, England, 2000.

- (50) Schautz, F.; Filippi, C. *J. Chem. Phys.* **2004**, *120*, 10931–10941.
- (51) Schautz, F.; Buda, F.; Filippi, C. *J. Chem. Phys.* **2004**, *121*, 5836–5844.
- (52) Cordova, F.; Doriol, L. J.; Ipatov, A.; Casida, M. E.; Filippi, C.; Vela, A. *J. Chem. Phys.* **2007**, *127*, 164111–1–16.
- (53) Tapavicza, E.; Tavernelli, I.; Röthlisberger, U.; Filippi, C.; Casida, M. E. *J. Chem. Phys.* **2008**, *129*, 124108–1–18.
- (54) Umrigar, C. J.; Toulouse, J.; Filippi, C.; Sorella, S.; Henning, R. G. *Phys. Rev. Lett.* **2007**, *98*, 110201–1–4.
- (55) Nightingale, M. P.; Melik-Alaverdian, V. *Phys. Rev. Lett.* **2001**, *87*, 043041–1–4.
- (56) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision C.02*; Gaussian, Inc.: Wallingford CT, 2004.
- (57) ADF2007 01, SCM, Theoretical Chemistry, Vrije Universiteit, Amsterdam, The Netherlands. <http://www.scm.com> (accessed June 15, 2009).
- (58) Fonseca Guerra, C.; Snijders, J. G.; te Velde, G.; Baerends, E. J. *Theor. Chem. Acc.* **1998**, *99*, 391–403.
- (59) te Velde, G.; Bickelhaupt, F. M.; van Gisbergen, S. J. A.; Fonseca Guerra, C.; Baerends, E. J.; Snijders, J. G.; Ziegler, T. *J. Comput. Chem.* **2001**, *22*, 931–967.
- (60) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100. Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- (61) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (62) Schipper, P. R. T.; Gritsenko, O. V.; van Gisbergen, S. J. A.; Baerends, E. J. *J. Chem. Phys.* **2000**, *112*, 1344–1352.
- (63) Karlström, G.; Lindh, R.; Malmqvist, P.-Å.; Roos, B. O.; Ryde, U.; Veryazov, V.; Widmark, P.-O.; Cossi, M.; Schimmelpfennig, B.; Neogrady, P.; Seijo, L. *Comput. Mater. Sci.* **2003**, *28*, 222–239.
- (64) Finley, J.; Malmqvist, P.-A.; Roos, B. O.; Serrano-Andrés, L. *Chem. Phys. Lett.* **1998**, *288*, 299–306.
- (65) Ghigo, C.; Roos, B. O.; Malmqvist, P.-A. *Chem. Phys. Lett.* **2004**, *396*, 142–149.
- (66) Roos, B. O.; Andersson, K. *Chem. Phys. Lett.* **1995**, *245*, 215–223.
- (67) MOLPRO, version 2006.1, a package of ab initio programs written by H.- J. Werner, P. J. Knowles, R. Lindh, F. R. Manby, M. Schütz, and others (Cardiff, UK, 2006). <http://www.molpro.net> (accessed June 15, 2009).
- (68) CHAMP is a quantum Monte Carlo program package written by C. J. Umrigar, C. Filippi, and collaborators.
- (69) Burkatzki, M.; Filippi, C.; Dolg, M. *J. Chem. Phys.* **2007**, *126*, 234105–1–8.
- (70) Filippi, C.; Umrigar, C. J. *J. Chem. Phys.* **1996**, *105*, 213–226. As the Jastrow correlation factor, we use the exponential of the sum of three fifth-order polynomials of the electron-nuclear (e-n), the electron-electron (e-e), and of pure 3-body mixed e-e and e-n distances, respectively. The Jastrow factor is adapted to deal with pseudo-atoms, and the scaling factor κ is set to 0.60 a.u.
- (71) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347–1363.
- (72) Casula, M. *Phys. Rev. B* **2006**, *74*, 161102(R)-1–4.
- (73) Shimomura, O. *FEBS Lett.* **1979**, *104*, 220–222.
- (74) Casida, M. E.; Casida, K. C.; Salahub, D. R. *Int. J. Quantum Chem.* **1998**, *70*, 933–941.
- (75) Wasserman, A.; Maitra, N. T.; Burke, K. *Phys. Rev. Lett.* **2003**, *91*, 263001-1–4.
- (76) Kamlet, M. J.; Abboud, J.-L. M.; Abraham, M. H.; Taft, R. R. *J. Org. Chem.* **1983**, *48*, 2877–2887.

CT900227J

JCTC

Journal of Chemical Theory and Computation

Effects of Restrained Sampling Space and Nonplanar Amino Groups on Free-Energy Predictions for RNA with Imino and Sheared Tandem GA Base Pairs Flanked by GC, CG, iGiC or iCiG Base Pairs

Ilyas Yildirim,[†] Harry A. Stern,[‡] Jiri Sponer,[§] Nada Spackova,[§] and Douglas H. Turner^{*‡}

Department of Physics and Astronomy, Department of Chemistry and Department of Pediatrics, University of Rochester, Rochester, New York, 14627, and Institute of Biophysics, Academy of Sciences of the Czech Republic, v.v.i., Královopolská 135, 61265 Brno, Czech Republic

Received December 5, 2008

Abstract: Guanine-adenine (GA) base pairs play important roles in determining the structure, dynamics, and stability of RNA. In RNA internal loops, GA base pairs often occur in tandem arrangements and their structure is context and sequence dependent. Calculations reported here test the thermodynamic integration (TI) approach with the amber99 force field by comparing computational predictions of free energy differences with the free energy differences expected on the basis of NMR determined structures of the RNA motifs (5'-GCGGACGC-3')₂, (5'-GCiGGAiCGC-3')₂, (5'-GGCGAGCC-3')₂, and (5'-GGiCGAiGCC-3')₂. Here, iG and iC denote isoguanosine and isocytidine, which have amino and carbonyl groups transposed relative to guanosine and cytidine. The NMR structures show that the GA base pairs adopt either imino (*cis* Watson–Crick/Watson–Crick A-G) or sheared (*trans* Hoogsteen/Sugar edge A-G) conformations depending on the identity and orientation of the adjacent base pair. A new mixing function for the TI method is developed that allows alchemical transitions in which atoms can disappear in both the initial and final states. Unrestrained calculations gave ΔG° values 2–4 kcal/mol different from expectations based on NMR data. Restraining the structures with hydrogen bond restraints did not improve the predictions. Agreement with NMR data was improved by 0.7 to 1.5 kcal/mol, however, when structures were restrained with weak positional restraints to sample around the experimentally determined NMR structures. The amber99 force field was modified to partially include pyramidalization effects of the unpaired amino group of guanosine in imino GA base pairs. This provided little or no improvement in comparisons with experiment. The marginal improvement is observed when the structure has potential cross-strand out-of-plane hydrogen bonding with the G amino group. The calculations using positional restraints and a nonplanar amino group reproduce the signs of ΔG° from the experimental results and are, thus, capable of providing useful qualitative insights complementing the NMR experiments. Decomposition of the terms in the calculations reveals that the dominant terms are from electrostatic and interstrand interactions other than hydrogen bonds in the base pairs. The results suggest that a better description of the backbone is key to reproducing the experimental free energy results with computational free energy predictions.

1. Introduction

Agreement between experiments and computational predictions is a test of our understanding of intermolecular

interactions. Molecular dynamics (MD) calculations are commonly used to provide insights into biological processes, including folding and dynamics of RNA.^{1–4} It is now possible to calculate MD trajectories with durations ap-

* Corresponding author. Phone: (585) 275-3207. Fax: (585) 276-0205. E-mail: turner@chem.rochester.edu.

[†] Department of Physics and Astronomy, University of Rochester.

[‡] Department of Chemistry, University of Rochester.

[§] Academy of Sciences of the Czech Republic.

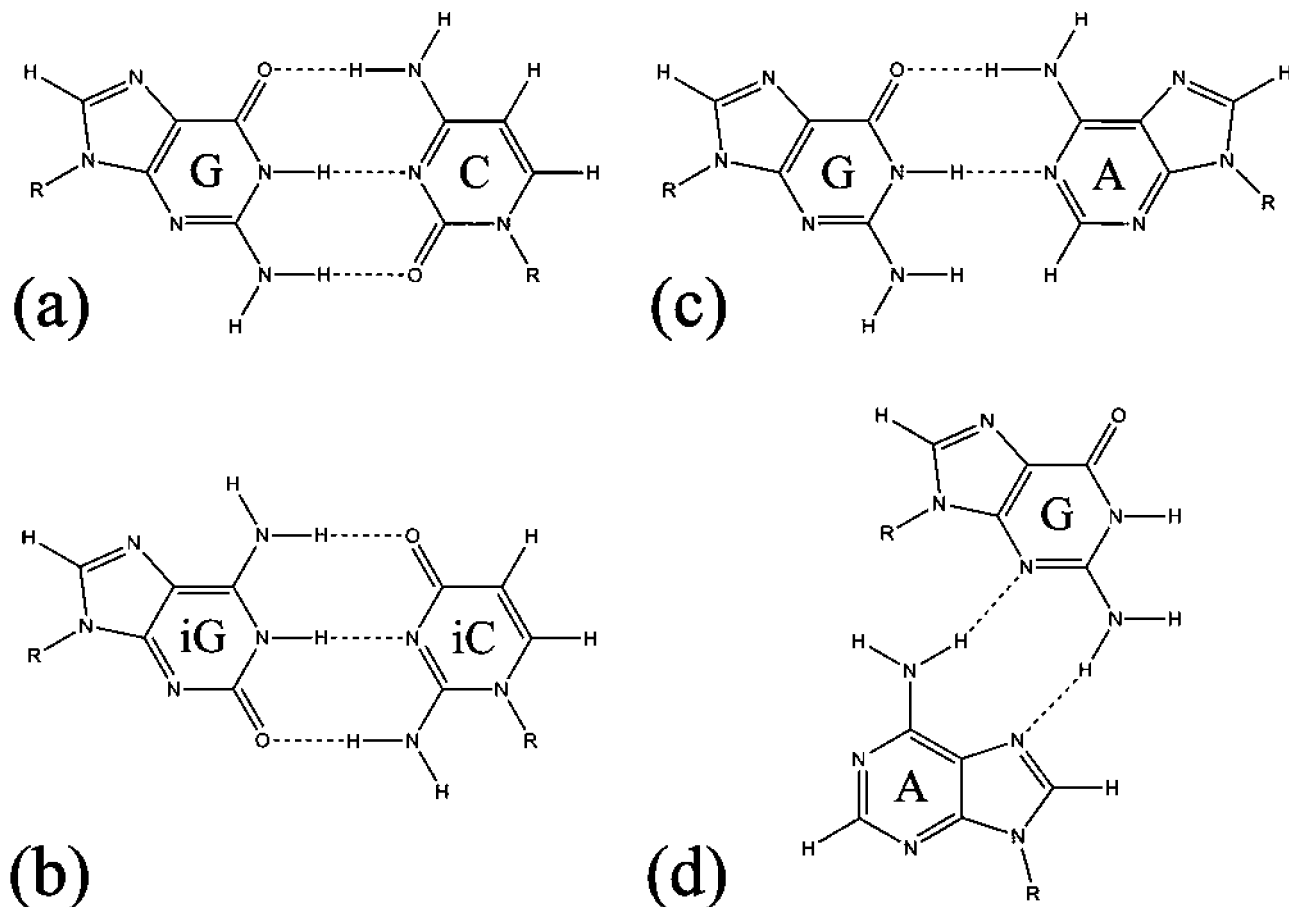


Figure 1. Structures of (a) Watson–Crick GC, (b) Watson–Crick iGiC, (c) Imino GA (*cis* Watson–Crick/Watson–Crick A-G), and (d) sheared GA (*trans* Hoogsteen/Sugar edge A-G). The difference of iG and iC from G and C, respectively, is that the amino and carbonyl groups are transposed. The amino group of G in an imino GA base pair is nonplanar,^{31,32} and it has the potential to form an extra cross-strand out-of-plane hydrogen bond.

proaching the millisecond time scale.^{5–9} These simulations, however, use approximate molecular mechanics (MM) potential energy functions or force fields,^{10–12} which are not expected to reproduce the true potential energy surface as well as quantum mechanical (QM) methods.

Nucleic acid structures are stabilized by a range of molecular interactions such as base stacking, base pairing, solvation, and ionic effects, but the balance of these interactions is complex and poorly understood.¹³ The structures and energetics of GA pairs in RNA oligonucleotides provide a particularly interesting case for testing computational methods. GA pairs occur in many RNAs.^{14–17} They have roles in tertiary interactions,^{18–22} metal binding,¹⁸ and protein recognition.²³ Often, GA pairs occur in tandem.^{14,24} NMR experiments^{25,26} reveal that these tandem GA pairs have either imino (*cis* Watson–Crick/Watson–Crick A-G) or sheared (*trans* Hoogsteen/Sugar edge A-G) conformations (Figure 1) depending on the adjacent base pairs, and this has a dramatic effect on global 3D structure (Figure 2). The interactions responsible for this sequence dependence are not fully understood.^{27,28} Thus, the sequence dependence of GA structures provides a benchmark for testing computational approaches.

Here, the sequence dependence of hydrogen-bonding patterns in GA pairs flanked by GC, CG, iGiC, or iCiG pairs is investigated with thermodynamic integration (TI)²⁹ cal-

culations, which may be used to estimate free energy differences. NMR structures^{25,26,30} show that the hydrogen-bonding patterns in tandem GA pairs change when the amino and carbonyl groups on flanking GC pairs are transposed to give flanking iGiC pairs (Figures 1 and 2). We use explicit solvent MD combined with TI to calculate ΔG°_2 and ΔG°_3 for the alchemical steps in the thermodynamic cycles shown in Figures 3 and 4. In both cycles

$$\Delta G^{\circ}_1 + \Delta G^{\circ}_2 = \Delta G^{\circ}_3 + \Delta G^{\circ}_4 \quad (1a)$$

$$\Delta G^{\circ}_3 - \Delta G^{\circ}_2 = \Delta G^{\circ}_1 - \Delta G^{\circ}_4 \quad (1b)$$

The right-hand side of eq 1b, $\Delta G^{\circ}_1 - \Delta G^{\circ}_4$, is either positive (Figure 3) or negative (Figure 4) on the basis of structures determined by NMR.^{25,26,30} For example, in Figure 3, ΔG°_1 and ΔG°_4 are positive and negative, respectively, because the NMR structures have imino and sheared GA pairs, respectively.^{26,30} Thus, the sign calculated by TI for $\Delta G^{\circ}_3 - \Delta G^{\circ}_2$ in eq 1b can be compared to the positive sign determined by experiment for $\Delta G^{\circ}_1 - \Delta G^{\circ}_4$ in eq 1b. When the standard TI procedure with standard amber99 force field was applied, the free energy calculations did not reproduce the expected sign for $\Delta G^{\circ}_3 - \Delta G^{\circ}_2$ in the cycle of Figure 3. Thus, the protocol was modified to test two hypotheses to explain this lack of agreement: (1) the amino group of G

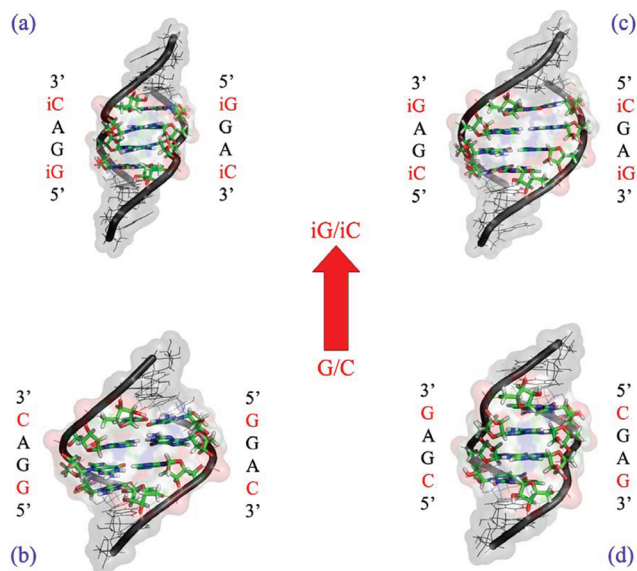


Figure 2. Comparison of structural motifs (a) $(5'(\underline{iGGAiC3'})_2)$ with (b) $(5'(\underline{GGAC3'})_2)$ and (c) $(5'(\underline{iCGAiG3'})_2)$ with (d) $(5'(\underline{CGAG3'})_2)$. Colored nucleotides in a–d are the loop regions of NMR structures of 2O81, 1MIS, 2O83, and 1YFV, respectively, from the PDB.^{25,26,30} Dramatic changes in the backbone are seen when the adjacent canonical base pairs of GC in $(5'(\underline{GGAC3'})_2)$ and CG in $(5'(\underline{CGAG3'})_2)$ are replaced with iGiC (b \rightarrow a) and iCiG (d \rightarrow c), respectively.

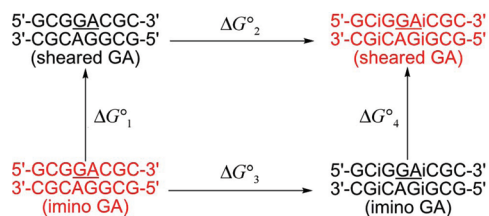


Figure 3. Thermodynamic cycle for $\underline{GGAC} \rightarrow \underline{iGGAiC}$ transformation. Structures in red are observed by NMR^{26,30} so that $\Delta G_1 - \Delta G_4$ is positive. PDB IDs of $(5'(\underline{GCGGACGC3'})_2)$ and $(5'(\underline{GCGGAiCGC3'})_2)$ are 1MIS and 2O81, respectively.

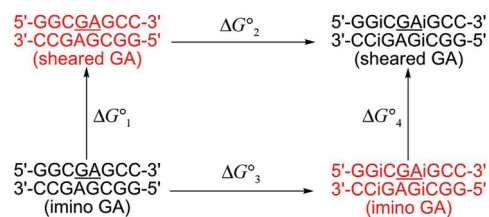


Figure 4. Thermodynamic cycle for $\underline{CGAG} \rightarrow \underline{iCGAiG}$ transformation. Structures in red are observed by NMR^{25,30} so that $\Delta G_1 - \Delta G_4$ is negative. PDB IDs of $(5'(\underline{GGCGAGCC3'})_2)$ and $(5'(\underline{GGiCGAiGCC3'})_2)$ are 1YFV and 2O83, respectively.

in an imino GA pair can be nonplanar,^{31–33} which is not considered by the force field, and (2) the conformational space sampled by MD simulations may include unrealistic conformations due to approximations in the force field.

The idea that the imino GA base pairs are involved in stabilizing out-of-plane molecular interactions via partial sp^3 hybridization was suggested by QM studies. High-level QM electronic structure calculations predict that GA base pairs

are intrinsically nonplanar in isolation, in contrast to canonical base pairs.^{31–33} Most striking is the nonplanarity of the *cis* Watson–Crick/Watson–Crick GA base pair (Figure 1c). In this base pair, the amino group of guanine is unpaired. In a planar base pair arrangement, it would face the H2 hydrogen of adenine, creating a repulsive contact. Instead, the guanine amino group adopts a substantially nonplanar (pyramidal) geometry due to partial sp^3 hybridization (see Supporting Information). This nonplanarity enables formation of weak out-of-plane hydrogen bonds by imino GA pairs. The nonplanarity is prevented by standard force fields. To test possible contributions of nonplanarity to determining structures of GA pairs, the intramolecular parameters for the amino group of the force field for guanine were modified and the effects of such modification were compared with results obtained with the standard force field.^{10,34} With this modification applied exclusively to those guanines involved in GA base pairs, the calculated sign of $\Delta G_3^\circ - \Delta G_2^\circ$ is consistent with the NMR experiments to within experimental error. Consideration of the nonplanar amino group suggests that the switch between imino and sheared GA base pairs is associated with the formation of an out-of-plane hydrogen bond between the G amino group and the flanking base pair.

The lack of agreement between computations and experiment could also reflect unphysical sampling of conformational space due to approximations in the force field. Thus, weak positional restraints were applied to limit the sampling to conformations resembling the NMR structures. The combination of positional restraints along with a nonplanar amino group of G in GA pairs provided the closest agreement between computational and experimental results.

2. Methods

2.1. Theory. The Helmholtz free energy difference between two systems, A and B, represented by Hamiltonians H_A and H_B , respectively, is given by

$$\Delta G^\circ = G_B - G_A = -kT \ln \langle e^{-\Delta H/k_B T} \rangle_A \quad (2)$$

Here $\Delta H = H_B - H_A$, T is temperature in kelvins, k_B is Boltzmann's constant, and $\langle \dots \rangle_A$ denotes the canonical ensemble average for system A.²⁹ One way to calculate this free energy difference, ΔG , is to use the TI approach.²⁹ This approach can be applied to unphysical alchemical transformations, such as changing a GC pair to an iGiC pair.

In the TI approach, a hybrid Hamiltonian $H(\lambda)$, which is a mix of both states A and B, is defined as

$$H(\lambda) = f_1(\lambda)H_A + f_2(\lambda)H_B \quad (3)$$

Here, $f_1(\lambda)$ and $f_2(\lambda)$ are mixing functions that depend on a mixing parameter, λ . The mixing parameter varies from 0 to 1 such that the hybrid Hamiltonian $H(\lambda) = H_A$ when $\lambda = 0$, and $H(\lambda) = H_B$ when $\lambda = 1$. For an effective calculation, it is desirable to use mixing functions that provide a smooth transformation from state A to state B.

With this approach, eq 2 can be rewritten as follows:

$$\Delta G^\circ = \int_0^1 \left(\frac{\partial H}{\partial \lambda} \right)_\lambda d\lambda \quad (4)$$

The transformation can be divided into discrete λ windows so that the integration in eq 4 may be performed numerically.

One choice of mixing functions, which is implemented in AMBER version 9,³⁵ is

$$\begin{aligned} f_1(\lambda, k) &= (1 - \lambda)^k \\ f_2(\lambda, k) &= 1 - (1 - \lambda)^k \end{aligned} \quad (5)$$

where k is an integer satisfying $k > 0$.³⁵ With these mixing functions, however, “dummy” atoms can only be used in the final state, B.³⁵ With this mixing rule, if dummy atoms are used in the initial state, A, then there is a convergence problem around $\lambda = 0$,³⁵ which does not allow a smooth transformation from A to B, thus precluding numerical integration to calculate ΔG in eq 4.^{36–38}

A new mixing rule was derived to allow dummy atoms in both the initial and final states. Let the mixing functions, $f_1(\lambda, k)$ and $f_2(\lambda, k)$, be defined as follows:

$$\begin{aligned} f_1(\lambda, k) &= f(\lambda, k) \\ f_2(\lambda, k) &= 1 - f(\lambda, k) \end{aligned} \quad (6)$$

satisfying the following conditions:

$$\begin{aligned} f(\lambda = 0, k) &= 1 \\ f(\lambda = 1, k) &= 0 \\ f^{(n)}(\lambda = 0, k) &= f^{(n)}(\lambda = 1, k) = 0 \end{aligned} \quad (7)$$

Here k is a positive integer ($k > 0$), $f^{(n)}$ is the n th derivative of the mixing function $f(\lambda, k)$, and $n < k$. Assuming that $f(\lambda, k)$ is a $(2k - 1)$ th-order polynomial in λ , the following mixing function can be derived with the help of Mathematica:³⁹

$$f(\lambda, k) = (1 - \lambda)^k \sum_{i=0}^{k-1} \frac{(k - 1 + i)!}{i!(k - 1)!} \lambda^i \quad (8)$$

This new mixing rule makes it possible to smoothly transform an initial state, A, to a final state, B, when dummy atoms are present in initial, final, or both states (Figure 5). We implemented this new mixing function in AMBER version 9.³⁵ Other methods have been proposed to deal with convergence problems arising from alchemical transformations involving the appearance or disappearance of atoms with hard-core Lennard-Jones potentials.^{40,41}

At each step of the simulation, the derivative of the hybrid Hamiltonian, $H(\lambda)$, with respect to the mixing parameter, λ , is calculated

$$\frac{\partial H(\lambda)}{\partial \lambda} = \frac{\partial f(\lambda)}{\partial \lambda} (H_A - H_B) \quad (9)$$

The derivatives of the old and new mixing functions are shown in Figure 5. The new mixing function improves convergence around $\lambda = 0$.

2.2. RESP Charge Calculation for iG, iC, G, A, C, and U. The systems in Figures 3 and 4 contain the unnatural nucleotides iC and iG. In order to be consistent and complete, the restrained electrostatic potential (RESP) charges were

derived for iG, iC, G, A, C, and U following the RESP protocol.^{10,42–44} These RESP charges were used in all the simulations. The charges for G, A, C, and U are similar to those in amber99.¹⁰

Dimethylphosphate (DMP) and the nucleosides of iG, iC, G, A, C, and U with C3'-endo sugar pucker were created with the xleap program in the AMBER package.³⁵ The molecules were optimized and the electrostatic potential at a set of gridpoints were calculated at the HF level using the 6-31G* basis set. These calculations were performed with Gaussian03.⁴⁵ RESP charges for these nucleosides were then calculated with the RESP program.⁴⁴ The only difference between these RESP calculations and those of Cornell et al. (1995)¹⁰ is that the sugar atoms were not made equivalent. That is, each of the nucleosides has different charges for the same sugar atoms.

Except for the RESP charges, the missing parameters for iG and iC were taken from the amber99 parameter set by analogy. The only difference of iG and iC from G and C is the transposition of the carbonyl and amino groups. In AMBER, this transposition does not change the type of carbonyl and amino group. Therefore, the missing bonded parameters of both the carbonyl and amino groups of iG and iC were taken to be identical to the parameters of G and C. As a result, this choice should not impact the final results.

A library of 18 residues was created (RX, RX3, and RX5 where $X = A, U, G, C, iG, \text{ and } iC$, and 3 and 5 denote the 3' and 5' end version of residue RX, respectively). Dummy atom residues for G, C, iG, and iC (RGD, RCD, IGD, ICD) were also created to allow alchemical transformations from G to iG and C to iC. The order of the atoms of either RGD and IGD or RCD and ICD were kept the same, which is mandatory for the TI approach in AMBER version 9 (see Supporting Information).

2.3. Explicit Solvent Simulations. For each $A \rightarrow B$ alchemical transformation (Figures 3 and 4), the NMR structures^{25,26,30} were used as the starting structures. These NMR structures are highlighted in red in Figures 3 and 4, and the 3D structures are shown in Figure 2. The structures were solvated with TIP3P water molecules⁴⁶ in an octahedral box with a buffer of 10 Å, which gave more than 3100 water molecules around the duplex. In each $A \rightarrow B$ alchemical transformation, states A and B had the same number of water molecules and box sizes. The system was neutralized with 14 Na⁺ ions. The parameter/topology files for the $A \rightarrow B$ transformations were created with the xleap module.³⁵

2.4. Force Fields. Two parameter sets were tested for each thermodynamic cycle: (1) the amber99 force field and (2) the amber99 force field with a modified frcmod file³⁵ (see Supporting Information), which allows nonplanarity of the amino group of G of the tandem GA base pairs.

Note that the modified force field does not fully capture the true energetics associated with amino group pyramidalization. Thus, the energies associated with amino group nonplanarity are most likely underestimated in this paper. Note also that when mimicking the pyramidalization effects for the purpose of condensed phase nucleic acid simulations, it is not advisable to fit the force field tightly to the gas-phase quantum chemical data (see Supporting Information).⁴⁷

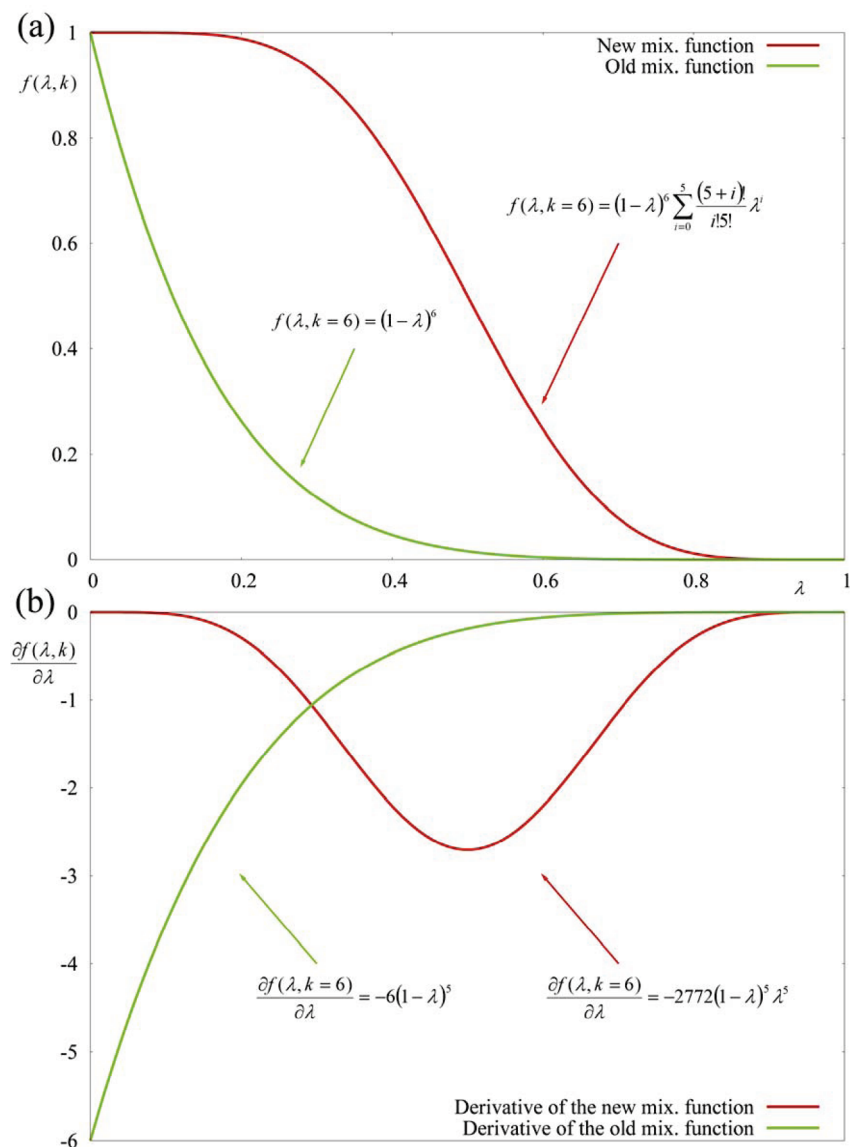


Figure 5. (a) Old (green) and new (red) mixing functions and (b) the derivatives of the old (green) and new (red) mixing functions for $k = 6$. In a simulation, eq 9 is calculated at each step, which depends on the derivative of the mixing function with respect to λ , $\partial f(\lambda, k)/\partial \lambda$. When dummy atoms are present in the initial state, there is a convergence problem around $\lambda = 0$ according to the old mixing rule, seen in (b).³⁵ The new mixing rule improves the convergence around $\lambda = 0$ by having a symmetric $\partial f(\lambda, k)/\partial \lambda$, as seen in (b). The plot was created with gnuplot (<http://www.gnuplot.info/>).

We also do not suggest using this modification for canonical base pairs where the amino group is planarized by planar hydrogen bonds.

2.5. Minimization. The structures were minimized in two steps. For each system, the same protocol was used: (1) With the RNA held fixed with a restraint force of 500 kcal/mol-Å², steepest descent minimization of 1000 steps was followed by a conjugate gradient minimization of 1500 steps. (2) With all restraints removed, steepest descent minimization of 1000 steps was followed by a conjugate gradient minimization of 1500 steps. During the minimization, an 8.0 Å long-range cutoff for nonbonded interactions was chosen.

2.6. Pressure Regulation. After the minimization, two steps of pressure equilibration were done on each system: (1) RNA structures were held fixed with a restraint force of 10 kcal/mol-Å². Constant volume dynamics with a long-range cutoff of 8.0 Å for nonbonded interactions was used.

SHAKE⁴⁸ was turned on for bonds involving hydrogen atoms, except for the amino hydrogen and dummy atoms of GC and iGiC base pairs flanking the tandem GA base pairs. The temperature was raised from 0 to 300 K in 20 ps. Langevin dynamics with a collision frequency of 1 ps⁻¹ was used. A total of 20 ps of MD were run with a 2 fs time step. (2) The same conditions as above were chosen, except that constant pressure dynamics with isotropic position scaling was turned on. Reference pressure was set to 1 atm with a pressure relaxation time of 2 ps. A total of 100 ps of MD were run with a 2 fs time step. The particle mesh Ewald (PME) method was on for all calculations (see Supporting Information).

Pressure relaxation decreases the box size of the system to bring the pressure to about 1 atm. The final restart file was used as the initial coordinate file for the λ simulations.

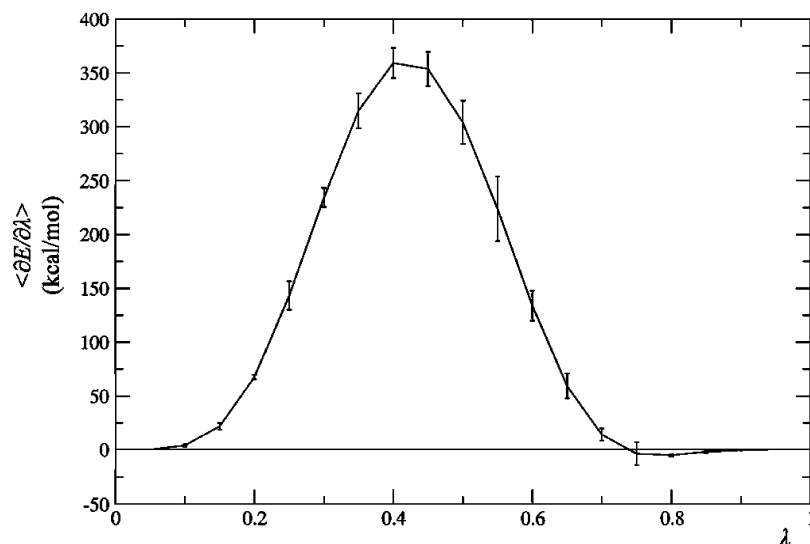


Figure 6. $\langle \partial E / \partial \lambda \rangle_\lambda$ as a function of λ for the $\text{CGAG} \rightarrow \text{iCGAiG}$ (sheared GA) transformation (unrestrained). For each λ simulation, the group-averaging method was used for error estimates by dividing each simulation into three parts. The plotted error bars, which are standard error of the mean, are enlarged 30-fold. The plot was created with GRACE (<http://plasma-gate.weizmann.ac.il/Grace/>).

2.7. λ Simulations. A total of 19 λ values were used, $\lambda = 0.05$ to $\lambda = 0.95$, with an increment of 0.05. The new mixing rule (with $k = 6$) was used in all λ simulations. For each λ simulation, the last structure of pressure regulation was taken as the initial structure. First, the structures were minimized following the minimization protocol described above. The production run was similar to the first step of the pressure equilibration described above. Constant volume dynamics was chosen with a long-range cutoff of 8.0 Å for nonbonded interactions. SHAKE was turned on for bonds involving hydrogen atoms, except the amino hydrogen and dummy atoms of GC and iGiC flanking the tandem GA base pairs. A total of 1 ns of MD was run at 300 K with a 1 fs time step. Another 1×10^6 time steps for $\lambda = 0.2$ to 0.8 values were run with a 1 fs time step, yielding a total of 2 ns of MD for these particular λ values. Because the TI calculations are done at constant volume, the calculated free energy changes are Helmholtz free energy changes. Because the PV contribution is small in the condensed phase, the difference between Helmholtz and Gibbs free energy differences is expected to be small.

2.8. Explicit Solvent Simulations with Positional Restraints. The same methodology described above was used with positional restraints to calculate each $A \rightarrow B$ alchemical transformation (Figures 3 and 4) for both the original amber99 and modified amber99 force fields. Harmonic restraints³⁵ with a force constant of 0.1 kcal/mol·Å² were applied to the backbone heavy atoms and to all atoms of the first two and last two base pairs of the structures. A total of 500 ps of MD was run with a 1 fs time step for all λ simulations. Another 5×10^5 time steps for $\lambda = 0.15$ to 0.85 of $\text{CGAG} \rightarrow \text{iCGAiG}$ (ΔG°_3 , imino) transformations were run with a 1 fs time step to attain better convergence, which yielded a total of 1 ns of MD for these particular λ values (see Supporting Information).

2.9. Explicit Solvent Simulations with Hydrogen Bond Restraints. The same methodology described above was used with distance restraints on hydrogen bonds to

calculate each $A \rightarrow B$ alchemical transformation (Figures 3 and 4) using both the original amber99 and modified amber99 force fields. Distance restraints of 25 kcal/mol·Å² were applied to all base pair hydrogen bonds (see Supporting Information). A total of 1 ns of MD was run with a 1 fs time step for all λ simulations. Another 5×10^5 MD steps for $\lambda = 0.15$ to 0.85 of $\text{CGAG} \rightarrow \text{iCGAiG}$ (ΔG°_2 , sheared, amber99), $\text{CGAG} \rightarrow \text{iCGAiG}$ (ΔG°_3 , imino, modified amber99), and $\text{GGAC} \rightarrow \text{iGGAiC}$ (ΔG°_3 , imino, modified amber99) transformations were run with a 1 fs time step to attain better convergence, which yielded a total of 1.5 ns of MD for these particular λ values (see Supporting Information).

2.10. Analysis. The first 250 ps of each λ simulation were omitted from the calculations to allow equilibration. For each λ simulation, $\langle \partial E / \partial \lambda \rangle_\lambda$ was calculated, where E is potential energy. The group-averaging method was used for error estimates by dividing each λ simulation into three parts. Error estimates are the standard error of the mean.

The trapezoidal rule was used to numerically integrate ΔG° in eq 4. A typical λ vs $\langle \partial E / \partial \lambda \rangle_\lambda$ plot is shown in Figure 6. For calculation of the error in ΔG° , $\langle \partial E / \partial \lambda \rangle_\lambda$ values were assumed to be independent.

Rmsd calculations were done with the ptraj module of AMBER 9.³⁵ Water molecules and Na⁺ ions were stripped out of the trajectory files, and all atoms of the RNA were included in the rmsd calculations. Each snapshot in the trajectory file was aligned with the initial starting structure (see Supporting Information for a typical time vs rmsd plot). For unrestrained simulations, the rmsd typically fluctuates between 1 and 2 Å, which implies that the simulation is sampling configurations close to the initial structure. For one or two unrestrained λ simulations, however, configurations with rmsd values greater than 3 Å are sampled. For the λ simulations with positional restraints, rmsd values are around 0.7 Å, which implies that only configurations very close to the initial structure are sampled. For the λ simulations with

Table 1. TI Method with Unrestrained and Restrained Molecular Dynamics with (or in parentheses without) Nonplanar G Amino Group in GA Base Pairs^{a,b}

alchemical transformation	ΔG°_2	ΔG°_3	$\Delta G^\circ_3 - \Delta G^\circ_2$	$\Delta G^\circ_1 - \Delta G^\circ_4$ (experimental)
No Restraints				
GGAC \rightarrow iGGAiC	111.227 \pm 0.100	111.155 \pm 0.110	-0.1 \pm 0.2 (-0.7 \pm 0.1)	\geq 3.1
CGAG \rightarrow iCGAiG	111.306 \pm 0.086	110.755 \pm 0.073	-0.6 \pm 0.1 (-0.6 \pm 0.1)	\leq -2.2
Positional Restraints				
GGAC \rightarrow iGGAiC	109.96 \pm 0.16	110.45 \pm 0.15	0.5 \pm 0.2 (0.0 \pm 0.2)	\geq 3.1
CGAG \rightarrow iCGAiG	112.23 \pm 0.18	110.41 \pm 0.11	-1.8 \pm 0.2 (-2.1 \pm 0.2)	\leq -2.2
Hydrogen Bond Restraints				
GGAC \rightarrow iGGAiC	111.31 \pm 0.10	111.30 \pm 0.07	0.0 \pm 0.1 (-1.0 \pm 0.2)	\geq 3.1
CGAG \rightarrow iCGAiG	111.15 \pm 0.10	111.12 \pm 0.09	0.0 \pm 0.1 (-0.6 \pm 0.1)	\leq -2.2

^a Results in kcal/mol. ^b See Supporting Information for details.

distance restraints on hydrogen bonds, rmsd values are around 1.5–2.0 Å.

2.11. Energy Decomposition. Free energy, ΔG° , can be decomposed into contributions resulting from bond, angle, dihedral, electrostatic, and van der Waals interactions:

$$\Delta G^\circ = \Delta G^\circ_{\text{bond}} + \Delta G^\circ_{\text{angle}} + \Delta G^\circ_{\text{dihedral}} + \Delta G^\circ_{\text{es}} + \Delta G^\circ_{\text{vdw}} \quad (10a)$$

where

$$\Delta G^\circ_{\text{bond}} = \int_0^1 \left\langle \frac{\partial E_{\text{bond}}}{\partial \lambda} \right\rangle_{\lambda} d\lambda \quad (10b)$$

etc.

Another way of decomposing the free energy is

$$\Delta G^\circ = \Delta G^\circ_{\text{RNA-env}} + \Delta G^\circ_{\text{HB}} + \Delta G^\circ_{\text{cross}} + \Delta G^\circ_{\text{ss}} + \Delta G^\circ_{\text{other}} \quad (11)$$

where $\Delta G^\circ_{\text{RNA-env}}$, $\Delta G^\circ_{\text{HB}}$, $\Delta G^\circ_{\text{cross}}$, and $\Delta G^\circ_{\text{ss}}$ represent the RNA-environment, hydrogen bonds within base pairs, cross-strand stacking, and single-strand stacking, respectively. $\Delta G^\circ_{\text{other}}$ includes the alchemical transformations of the individual bases. Both types of free energy decompositions were done on the alchemical transformations with positional restraints, ΔG°_2 and ΔG°_3 in Figures 3 and 4, to examine qualitatively the dominant factors in the TI calculations.

For decompositions, the trajectory files of each restrained λ simulation were used. The first 250 ps of the trajectories were omitted. The remaining structures generated every 1 ps were used to calculate the individual terms in eqs 10a and 11 using the TI approach. Details of the decompositions are described in Supporting Information.

Because the decompositions were done on alchemical transformations, they do not have a direct physical meaning. Comparisons of $(\Delta G^\circ_3 - \Delta G^\circ_2)$ values, however, give an idea of the contributions of individual terms.

3. Results

Two thermodynamic cycles are analyzed, (GCGGACGC)₂ \rightarrow (GCiGGAiCGC)₂ (Figure 3) and (GGCGAGCC)₂ \rightarrow (GGiCGAiGCC)₂ (Figure 4). The TI approach was used to calculate ΔG°_2 and ΔG°_3 in eq 1a to see if the amber99 force field or its modified version, that allows a nonplanar amino group of G in GA pairs, is consistent with the experimental

results for $\Delta G^\circ_1 - \Delta G^\circ_4$ (see eq 1b). For both unrestrained and restrained simulations, convergence analysis shows that the free energies are well converged (see Supporting Information).

3.1. GGAC \rightarrow iGGAiC Transformation (Unrestrained Simulations). NMR structures (Figure 2a and b) show that tandem GA base pairs flanked by GC have primarily imino GA base pairs,²⁶ while when flanked by iGiC, they have primarily sheared GA base pairs (Figure 3).³⁰ On the basis of the NMR spectra,^{26,30} there could be as much as 5% sheared GA for (GCGGACGC)₂ and 10% imino GA for (GCiGGAiCGC)₂. At 300 K, these limits amount to $\Delta G^\circ_1 \approx -R(300) \ln(1/19) = 1.8$ kcal/mol and $\Delta G^\circ_4 \approx -R(300) \ln(9) = -1.3$ kcal/mol. Smaller populations of the minor species would increase the magnitude of each ΔG° . Thus eq 1b becomes

$$\Delta G^\circ_3 - \Delta G^\circ_2 = \Delta G^\circ_1 - \Delta G^\circ_4 \geq 3.1 \text{ kcal/mol} \quad (12)$$

Table 1 shows the results of the TI calculations for the GGAC \rightarrow iGGAiC transformation when unrestrained (see also Supporting Information). The amber99 force field gives a free energy difference of -0.7 ± 0.1 kcal/mol for $\Delta G^\circ_3 - \Delta G^\circ_2$, while the modified amber99 force field yields a free energy difference of -0.1 ± 0.2 kcal/mol. Neither force field predicts the experimental result of ≥ 3.1 kcal/mol, but modified amber99 improves the predictions by 0.6 kcal/mol.

In an imino GA base pair, the amino group of G is not hydrogen bonded with adenosine (Figure 1). As a result, this amino group favors a conformation³² where it can form an out-of-plane hydrogen bond.⁴⁹ When GA base pairs are in an imino conformation in the GGAC system, there are two potential cross-strand out-of-plane hydrogen bonds between the amino groups of G of GA base pairs with the carbonyl groups of C of the flanking Watson–Crick GC base pairs. The inclusion of nonplanarity of the amino group of G in an imino GA base pair strengthens the effects of this cross-strand interaction. Figure 7 shows the cross-strand distance of H21–O2 after minimization with (a) amber99 and (b) modified amber99. This distance is around 3.0 Å in amber99, while it is around 1.8 Å in modified amber99. As a result, this out-of-plane hydrogen bond is strengthened with modified amber99. The nonplanar amino group of G in an imino GA base pair is a quantum mechanical effect^{31,32} that improves the free energy prediction, but not by enough to

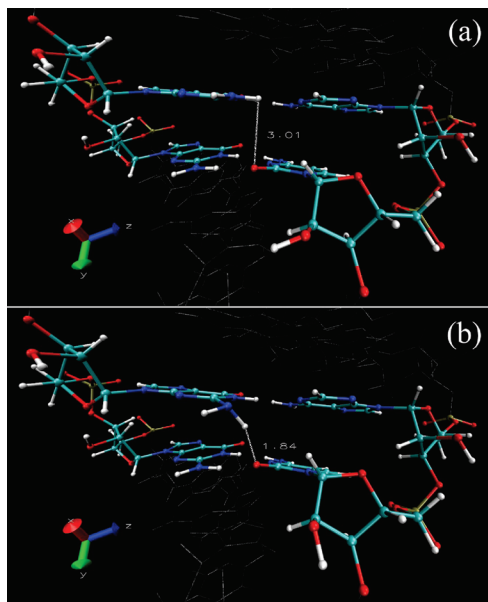


Figure 7. 5'GG3'/3'CA5', imino GA stacked on GC in 5'GCGGACGC3'/3'CGCAGGCG5' (1MIS) after minimization with (a) amber99 force field and (b) modified amber99 force field. The cross-strand hydrogen bond of H2–O2 distance is reduced from 3.01 to 1.84 Å when nonplanarity on amino group of G in a GA base pair is imposed on the structure. The figures were created with VMD.⁵⁶

be consistent with the experimental results. Evidently, neither amber99 nor the modified amber99 force fields accurately take into account all the interactions responsible for the observed structures.

3.2. CGAG → iCGAiG Transformation (Unrestrained Simulations). NMR structures (Figure 2c and d) show that tandem GA base pairs flanked by CG have primarily sheared GA base pairs,²⁵ while when flanked by iCiG, they have primarily imino GA base pairs (Figure 4).³⁰ On the basis of the NMR spectra, minor species are possible at roughly 5 and 33%, respectively, for (GGCGAGCC)₂ and (GGiCGAiGCC)₂, so $\Delta G^{\circ}_1 \approx -R(300) \ln(19/1) = -1.8$ kcal/mol and $\Delta G^{\circ}_4 \approx -R(300) \ln(1/2) = 0.4$ kcal/mol. Thus eq 1b becomes

$$\Delta G^{\circ}_3 - \Delta G^{\circ}_2 = \Delta G^{\circ}_1 - \Delta G^{\circ}_4 \leq -2.2 \text{ kcal/mol} \quad (13)$$

Both the amber99 and modified amber99 force fields give a free energy difference of -0.6 ± 0.1 kcal/mol for $\Delta G^{\circ}_3 - \Delta G^{\circ}_2$ (Table 1). In the iCGAiG system, the cross-strand distance between the amino group (H21) of G of the GA base pair and the carbonyl group (O4) of iC of the iCiG base pair is longer than in iGGAiC (compare Figures 7 and 8). As a result, the prediction of modified amber99 is similar to the prediction of amber99. The magnitude of $\Delta G^{\circ}_3 - \Delta G^{\circ}_2$ is smaller than suggested by experiments. Again, neither amber99 nor the modified amber99 force fields accurately take into account all the interactions responsible for the observed structures.

3.3. GGAC → iGGAiC Transformation (Simulations with Positional Restraints). Table 1 shows the results of the TI calculations with positional restraints. The amber99 and modified amber99 force fields, respectively, give free energy

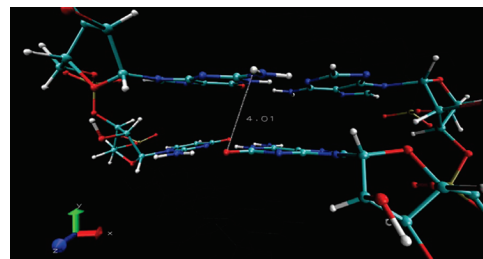


Figure 8. 5'iCG3'/3'iGA5', imino GA stacked on iCiG in 5'GGiCGAiGCC3'/3'CCiGAGiCGG5' (2O83) after minimization with amber99 force field. The distance between the cross-strand hydrogen bond of H21–O6 is 4.01 Å. With the modified amber99 force field, this distance is 2.60 Å. The figure was created with VMD.⁵⁶

differences, $\Delta G^{\circ}_3 - \Delta G^{\circ}_2$, of 0.0 ± 0.2 and 0.5 ± 0.2 kcal/mol for the GGAC → iGGAiC transformation. Thus, restraining the backbone gives a value about 0.6 kcal/mol closer to the lower limit of 3.1 kcal/mol for the experimental free energy difference (eq 12 and Table 1). The λ simulations are unphysical with hybrid Hamiltonians defined in eq 3. Some of the λ simulations of unrestrained TI calculations have rmsd values greater than 3 Å for more than half the time, meaning that most configurations sampled do not resemble the NMR structures. The largest rmsd values are reduced to about 0.7 Å when restraints are applied, which reduces sampling of conformations that do not resemble the NMR structures and, most likely, provides a smoother transformation from state A to state B. Evidently, forcing sampling around the experimentally determined structures gives better results.

Free energy decompositions of the TI calculations with positional restraints according to eq 10a show that the dominant contributions to the ΔG° for the alchemical transformations are due to electrostatics (Table 2). Decompositions according to eq 11 (Table 3) show that the cross-strand interactions, excluding the base pair hydrogen bonding, provide the largest contributions toward a free energy difference consistent with the experimental results. For calculations using both the original amber99 and modified amber99 force fields, $(\Delta G^{\circ}_3 - \Delta G^{\circ}_2)_{\text{cross}}$ for GGAC → iGGAiC is positive, 44.9 ± 0.3 and 51.7 ± 0.3 kcal/mol, respectively (Table 3). Thus, inclusion of nonplanar amino groups of G in GA base pairs gives a result more in line with experiment ($\Delta G^{\circ}_3 - \Delta G^{\circ}_2 = \Delta G^{\circ}_3 - \Delta G^{\circ}_2 \geq 3.1$ kcal/mol), which is expected due to the improved cross-strand interaction of the amino group of G of the GA base pair with the adjacent base pair's carbonyl group.

It is noteworthy that the RNA–environment interaction, $(\Delta G^{\circ}_3 - \Delta G^{\circ}_2)_{\text{RNA-env}}$, which includes the RNA–solvent and RNA–counterion interactions, has almost no effect on the free energy differences for GGAC → iGGAiC both in amber99 and modified amber99 force fields.

One interesting result of the decomposition of GGAC → iGGAiC transformations is that the base pair hydrogen bond interactions of $\Delta G^{\circ}_{3,\text{HB}}$ and $\Delta G^{\circ}_{2,\text{HB}}$ (Table 3) always favor the structures with tandem GA base pairs closed by iGiC base pairs (by -1.1 and -3.0 kcal/mol in amber99 force field and by -1.3 and -2.7 kcal/mol in modified amber99

Table 2. Decomposition of Free Energy Differences (kcal/mol) to Bond, Angle, Dihedral, Electrostatic, and van der Waals Terms from TI Method using Molecular Dynamics with Positional Restraints^a

alchemical transformation	ΔG°	$\Delta G^\circ_{\text{bond}}$	$\Delta G^\circ_{\text{angle}}$	$\Delta G^\circ_{\text{dihedral}}$	$\Delta G^\circ_{\text{es}}$	$\Delta G^\circ_{\text{vdw}}$
amber99						
GGAC \rightarrow iGGAiC (ΔG°_3 , imino GA)	110.2 \pm 0.2	-5.0 \pm 0.1	0.4 \pm 0.0	0.2 \pm 0.0	113.4 \pm 0.2	1.0 \pm 0.0
GGAC \rightarrow iGGAiC (ΔG°_2 , sheared GA)	110.0 \pm 0.3	-4.7 \pm 0.1	0.3 \pm 0.0	0.2 \pm 0.0	112.7 \pm 0.3	1.4 \pm 0.0
($\Delta G^\circ_3 - \Delta G^\circ_2$)	0.2 \pm 0.4	-0.3 \pm 0.1	0.1 \pm 0.0	0.0 \pm 0.0	0.7 \pm 0.4	-0.4 \pm 0.0
CGAG \rightarrow iCGAiG (ΔG°_3 , imino GA)	110.6 \pm 0.1	-5.2 \pm 0.1	0.6 \pm 0.0	0.2 \pm 0.0	114.3 \pm 0.1	0.7 \pm 0.0
CGAG \rightarrow iCGAiG (ΔG°_2 , sheared GA)	112.5 \pm 0.3	-5.4 \pm 0.2	0.4 \pm 0.0	0.2 \pm 0.0	115.9 \pm 0.3	1.4 \pm 0.0
($\Delta G^\circ_3 - \Delta G^\circ_2$)	-1.9 \pm 0.3	0.2 \pm 0.2	0.2 \pm 0.0	0.0 \pm 0.0	-1.6 \pm 0.3	-0.7 \pm 0.0
amber99 with nonplanar G amino group in GA base pairs						
GGAC \rightarrow iGGAiC (ΔG°_3 , imino GA)	110.6 \pm 0.3	-4.7 \pm 0.2	0.3 \pm 0.0	0.2 \pm 0.0	113.5 \pm 0.2	1.2 \pm 0.0
GGAC \rightarrow iGGAiC (ΔG°_2 , sheared GA)	110.0 \pm 0.3	-4.8 \pm 0.1	0.3 \pm 0.0	0.2 \pm 0.0	112.9 \pm 0.2	1.3 \pm 0.0
($\Delta G^\circ_3 - \Delta G^\circ_2$)	0.6 \pm 0.4	0.1 \pm 0.2	0.0 \pm 0.0	0.0 \pm 0.0	0.6 \pm 0.3	-0.1 \pm 0.0
CGAG \rightarrow iCGAiG (ΔG°_3 , imino GA)	110.8 \pm 0.1	-5.0 \pm 0.0	0.6 \pm 0.0	0.3 \pm 0.0	114.1 \pm 0.1	0.8 \pm 0.0
CGAG \rightarrow iCGAiG (ΔG°_2 , sheared GA)	111.9 \pm 0.2	-5.7 \pm 0.2	0.5 \pm 0.0	0.2 \pm 0.0	115.6 \pm 0.3	1.4 \pm 0.0
($\Delta G^\circ_3 - \Delta G^\circ_2$)	-1.1 \pm 0.2	0.7 \pm 0.2	0.1 \pm 0.0	0.1 \pm 0.0	-1.5 \pm 0.3	-0.6 \pm 0.0

^a The values for ΔG° differ somewhat from those in Table 1 because the data set for the table includes structures generated every 50 fs, while the data set for this table contained only structures generated every 1 ps.

Table 3. Decomposition of Free Energy Differences (kcal/mol) to RNA Environment, Hydrogen Bond, Cross-Strand Stacking, Single-Strand Stacking, and Other Interactions^{a,b}

alchemical transformation	ΔG°	$\Delta G^\circ_{\text{RNA-env}}$	$\Delta G^\circ_{\text{HB}}$	$\Delta G^\circ_{\text{cross}}$	$\Delta G^\circ_{\text{ss}}$	$\Delta G^\circ_{\text{other}}^c$
amber99						
GGAC \rightarrow iGGAiC (ΔG°_3 , imino GA)	109.5 \pm 0.5	1.7 \pm 0.3	-1.1 \pm 0.1	-10.5 \pm 0.2	-18.2 \pm 0.2	137.6 \pm 0.3
GGAC \rightarrow iGGAiC (ΔG°_2 , sheared GA)	109.6 \pm 0.6	1.7 \pm 0.4	-3.0 \pm 0.1	-55.4 \pm 0.2	27.1 \pm 0.3	139.2 \pm 0.3
($\Delta G^\circ_3 - \Delta G^\circ_2$)	-0.1 \pm 0.8	0.0 \pm 0.5	1.9 \pm 0.1	44.9 \pm 0.3	-45.3 \pm 0.4	-1.6 \pm 0.4
CGAG \rightarrow iCGAiG (ΔG°_3 , imino GA)	111.4 \pm 0.3	0.3 \pm 0.1	-1.8 \pm 0.1	-13.8 \pm 0.1	-9.6 \pm 0.1	136.3 \pm 0.2
CGAG \rightarrow iCGAiG (ΔG°_2 , sheared GA)	113.0 \pm 0.6	-8.3 \pm 0.4	0.2 \pm 0.1	16.5 \pm 0.2	-30.6 \pm 0.2	135.2 \pm 0.3
($\Delta G^\circ_3 - \Delta G^\circ_2$)	-1.6 \pm 0.7	8.6 \pm 0.4	-2.0 \pm 0.1	-30.3 \pm 0.2	21.0 \pm 0.2	1.1 \pm 0.4
amber99 with nonplanar G amino group in GA base pairs						
GGAC \rightarrow iGGAiC (ΔG°_3 , imino GA)	110.0 \pm 0.6	0.9 \pm 0.3	-1.3 \pm 0.1	-6.6 \pm 0.2	-20.8 \pm 0.1	137.9 \pm 0.4
GGAC \rightarrow iGGAiC (ΔG°_2 , sheared GA)	109.7 \pm 0.6	1.3 \pm 0.3	-2.7 \pm 0.1	-58.3 \pm 0.2	30.7 \pm 0.3	138.7 \pm 0.4
($\Delta G^\circ_3 - \Delta G^\circ_2$)	0.3 \pm 0.8	-0.4 \pm 0.4	1.4 \pm 0.1	51.7 \pm 0.3	-51.5 \pm 0.3	-0.8 \pm 0.6
CGAG \rightarrow iCGAiG (ΔG°_3 , imino GA)	111.7 \pm 0.3	-1.2 \pm 0.2	-2.1 \pm 0.1	-23.7 \pm 0.1	1.7 \pm 0.1	137.0 \pm 0.2
CGAG \rightarrow iCGAiG (ΔG°_2 , sheared GA)	112.6 \pm 0.5	-7.6 \pm 0.3	-0.2 \pm 0.1	13.4 \pm 0.2	-30.1 \pm 0.2	137.1 \pm 0.3
($\Delta G^\circ_3 - \Delta G^\circ_2$)	-0.9 \pm 0.6	6.4 \pm 0.4	-1.9 \pm 0.1	-37.1 \pm 0.2	31.8 \pm 0.2	-0.1 \pm 0.4

^a Other interactions include alchemical transformations of the individual bases from the TI method with MD using positional restraints.

^b The sum of the ΔG° 's for each transformation differ from those in Table 2 even though the same structures were used because the interactions are decomposed in a totally different way as described in Supporting Information. ^c See Supporting Information.

force field). The results of an Individual Nearest-Neighbor Hydrogen-Bonding (INN-HB) model analysis of thermodynamic data for duplexes with GC and iGiC base pairs predict that the hydrogen bonding in two iGiC pairs should favor duplex formation by $4(-0.24) = -0.96$ kcal/mol at 300 K relative to that from two GC pairs.⁵⁰ Here, -0.24 kcal/mol is the favorable free energy increment at 300 K "per terminal iG-iC", which is half the increment per internal iG-iC.

3.4. CGAG \rightarrow iCGAiG Transformation (Simulations with Positional Restraints). Restrained simulations for the CGAG \rightarrow iCGAiG transformation with positional restraints give free energy differences of -2.1 ± 0.2 and -1.8 ± 0.2 kcal/mol for $\Delta G^\circ_3 - \Delta G^\circ_2$ with the amber99 and modified amber99 force fields, respectively (Table 1). Both results are consistent with NMR experiments because $\Delta G^\circ_3 - \Delta G^\circ_2 \leq -2.2$ kcal/mol (eq 13 and Table 1). Rmsd values in the λ simulations are around 0.7 Å, similar to those obtained in the restrained GGAC \rightarrow iGGAiC transformation with positional restraints. Evidently, imposing restraints restricts sampling to conformations resembling the NMR structures in the TI calculations. Compared to the unrestrained TI calculation results for the CGAG \rightarrow iCGAiG transformation, the free energy value is closer to that expected from NMR

experiments by around 1 kcal/mol for both the original amber99 and modified amber99 force fields.

Similar to the GGAC \rightarrow iGGAiC transformation, the dominant terms contributing to the free energy differences for the CGAG \rightarrow iCGAiG transformation are due to electrostatics (Table 2) and cross-strand interactions (Table 3). Values for $(\Delta G^\circ_3 - \Delta G^\circ_2)_{\text{cross}}$ for amber99 and modified amber99 are -30.3 ± 0.2 and -37.1 ± 0.2 kcal/mol, respectively (Table 3), which provide the largest contribution toward a free energy difference consistent with the experimental results. Again, inclusion of nonplanar amino groups of G in GA base pairs results in cross-strand interactions more in line with experiment ($\Delta G^\circ_3 - \Delta G^\circ_2 = \Delta G^\circ_1 - \Delta G^\circ_4 \leq -2.2$ kcal/mol).

The contributions of the RNA-environment interaction, $(\Delta G^\circ_3 - \Delta G^\circ_2)_{\text{RNA-env}}$, are 8.6 ± 0.5 and 6.4 ± 0.4 kcal/mol in amber99 and modified amber99 force fields, respectively (Table 3). This result is different from that obtained for the GGAC \rightarrow iGGAiC transformation. The result suggests that the free energy contribution of the RNA-environment interaction disfavors the experimental results. Most of this contribution is due to the CGAG \rightarrow iCGAiG sheared GA transformation. The values of $\Delta G^\circ_{2,\text{RNA-env}}$ from am-

ber99 and modified amber99 calculations for the $\text{CGAG} \rightarrow \text{iCGAiG}$ alchemical transformations are -8.3 ± 0.4 and -7.6 ± 0.3 kcal/mol, respectively. Even though this quantity does not have any physical meaning, it implies that replacing the adjacent GC base pairs of the tandem sheared GA base pairs in $(5'\text{-GGCGAGCC-}3')_2$ with iGiC base pairs results in a more favorable RNA–environment interaction. This is probably due to the different electronic structures of GC and iGiC base pairs, which will interact with water and counterions differently.

Similar to the $\text{GGAC} \rightarrow \text{iGGAiC}$ transformation, the hydrogen-bonding interactions of $\Delta G_{3,\text{HB}}^\circ$ favor the structures with tandem GA base pairs closed by iCiG base pairs by -1.8 and -2.1 kcal/mol in amber99 and modified amber99 force fields, respectively, (Table 3) which is consistent with the INN-HB analysis of Chen et al. (2001).⁵⁰ Values of $\Delta G_{2,\text{HB}}^\circ$ from amber99 and modified amber99 force fields, however, are almost zero. The eight calculations give an average predicted enhancement of 1.5 kcal/mol in stability from substituting two GC pairs with iGiC pairs. This value is similar to the enhancement of 0.96 kcal/mol at 300 K expected on the basis of the INN-HB model.⁵⁰

3.5. $\text{GGAC} \rightarrow \text{iGGAiC}$ and $\text{CGAG} \rightarrow \text{iCGAiG}$ Transformation (Simulations with Hydrogen Bond Restraints). Table 1 shows the results of the restrained TI calculations with hydrogen bond restraints. The amber99 and modified amber99 force fields, respectively, give free energy differences, $\Delta G_3^\circ - \Delta G_2^\circ$, of -1.0 ± 0.2 and 0.0 ± 0.1 kcal/mol for the $\text{GGAC} \rightarrow \text{iGGAiC}$ transformation and -0.6 ± 0.1 and 0.0 ± 0.1 kcal/mol for the $\text{CGAG} \rightarrow \text{iCGAiG}$ transformation. Compared to the unrestrained simulations, there is no improvement in the free energy predictions. Yet, inclusion of nonplanarity in the amino groups of G in GA base pairs gives a free energy difference for the $\text{GGAC} \rightarrow \text{iGGAiC}$ transformation that is closer to the experimental value by 1 kcal/mol. TI calculations with hydrogen bond restraints imply that the sampling problem is due to the improper description of the backbone torsions.

4. Discussion

Non-Watson–Crick base pairs are common in RNA and each has more than one possible conformation.⁵¹ Predicting the dependence of these conformations on sequence context provides a test of force fields.²⁸ A new mixing function allows TI calculations that can be compared with NMR results for the structural change of GA base pairs associated with replacing adjacent GC pairs with iGiC pairs. Allowing nonplanar G amino groups in GA pairs and applying weak positional restraints to sample around NMR determined structures improves agreement with experiment by 1.2 kcal/mol for both systems studied and provides the same signs for the experimental and computational free energy results (Table 1). Distance restraints on hydrogen bonds alone or in combination with nonplanar G amino groups do not result in agreement between experiment and computations. This result suggests that the backbone torsions in the amber99 force field may allow sampling of unphysical conformations. The amber99 force field provides good predictions for structures of canonical base pairs.³⁴ The backbone torsions

of noncanonical base pairs (e.g., Figure 2), however, can be different from canonical base pairs. A better description of backbone torsions to include the characteristics of noncanonical base pairs might improve sampling.

While TI calculations give values for $\Delta G_3^\circ - \Delta G_2^\circ$ with the same sign as experimental values, they do not reproduce the magnitudes expected from experimental results (Table 1). The error estimates in Table 1 reflect the statistical uncertainties of the individual TI runs but do not include systematic errors resulting from choosing starting structures, structural rearrangements, substates, and relaxations that occur on time scales longer than the simulation runs. Also, they do not include errors due to the force field approximations (see Supporting Information). Given these uncertainties and the number and magnitudes of individual terms (Tables 1–3), even agreement between the calculated and experimental signs of free energy differences is notable (Table 1).

The inclusion of a nonplanar amino group of G in GA base pairs improves the predictions of free energy differences between sheared and imino GA pair conformations if there are potential out-of-plane hydrogen-bonding sites for the free amino group of G in an imino GA pair (Table 1 and Figures 7 and 8). The nonplanarity can strengthen this hydrogen bond by reducing the distance between the partial positive charge on the amino hydrogen and the partial negative charge on other atoms. Allowing nonplanarity has little effect on the free energy predictions of systems that have sheared GA base pairs because the amino group of G in a sheared GA base pair already forms a hydrogen bond with adenosine (Figure 1). A force field cannot fully capture the true electronic structure effects associated with the amino group's delicate sp^2 – sp^3 balance (see Supporting Information). The force field with its few parameters and constant atom-centered point charges merely allow the amino group to be sufficiently flexible to allow for nonplanarity. It cannot reflect the true flexibility of the amino group electronic structure upon continually changing the amino group geometry and environment, charge redistributions, lone pair formation, etc. Thus, the free energy effects of the nonplanarity are most likely underestimated. Even so, the force field with nonplanarity leads to a local structure consistent with expectations from crystal structures and sequence comparisons.³² Note that the modified force field was applied only for guanines involved in GA base pairs (see Supporting Information). For all other amino groups, standard parameters assuming purely planar amino groups were used, as amino groups in canonical base pairs exhibit sp^2 hybridization.

Additional interactions may also be important for accurately describing the sequence dependence of structures of GA pairs. For example, the current force fields use fixed atom-centered charges to mimic the electronic environment of the molecules. Polarizable force fields,^{52–55} when completely developed for nucleic acid simulations, with geometry-dependent electrostatics, including flexible lone pairs, might improve the free energy predictions for the types of alchemical transformations presented in this paper. Nevertheless, the calculations come close to reproducing the conformational trends revealed by NMR structures. Thus, such computations can provide qualitative insights to complement experiments.

Qualitative insights into the effects of different interactions are provided by the energy decompositions in Tables 2 and 3. The results in Table 2 show that the largest free energy term is due to electrostatic interactions. The second decomposition method (Table 3) shows that all the cross-strand interactions favor the NMR results. Moreover, replacement of GC with iGiC base pairs enhanced hydrogen bonding in the base pairs, consistent with the interpretation of experimental thermodynamic results.⁵⁰

Acknowledgment. We thank Prof. Scott Kennedy for help with Figure 2 and Dr. Daniel Svozil for helpful discussions. This work was supported by NIH grant GM22939 (D.H.T.), by the Academy of Sciences of the Czech Republic, grants no. AV0Z50040507 and AV0Z50040702, grant IAA400040802 by Grant Agency of the Academy of Sciences of the Czech Republic, grant LC06030, Ministry of Education of the Czech Republic and grant 203/09/1476, Grant Agency of the Czech Republic (J.S and N.S.). Preliminary calculations utilized the NCSA Xeon IA-32 Linux Cluster supported by the National Center for Supercomputing Applications under CHE060041T.

Supporting Information Available: Modified force field parameters for guanine with nonplanar amino group; description of the force field modification; a typical plot of rmsd as a function of time; modified force field parameters for isoguanosine, isocytidine and dummy atoms; RESP charges for C, iC, G, iG, A, and U; convergence analysis of the unrestrained alchemical transformations; decomposition of the free energies of the restrained alchemical transformations; details of individual TI calculations with unrestrained simulations of GGAC → iGG*A*iC; examples of parameters used in the λ simulations; example hydrogen bond restraint file; detailed version of Table 1. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Krasovska, M. V.; Sefcikova, J.; Spackova, N.; Sponer, J.; Walter, N. G. Structural dynamics of precursor and product of the RNA enzyme from the hepatitis delta virus as revealed by molecular dynamics simulations. *J. Mol. Biol.* **2005**, *351*, 731.
- (2) Csaszar, K.; Spackova, N.; Stefl, R.; Sponer, J.; Leontis, N. B. Molecular dynamics of the frame-shifting pseudoknot from beet western yellows virus: The role of non-Watson-Crick base-pairing, ordered hydration, cation binding and base mutations on stability and unfolding. *J. Mol. Biol.* **2001**, *313*, 1073.
- (3) Auffinger, P.; Hashem, Y. Nucleic acid solvation: from outside to insight. *Curr. Opin. Struct. Biol.* **2007**, *17*, 325.
- (4) McDowell, S. E.; Spackova, N.; Sponer, J.; Walter, N. G. Molecular dynamics simulations of RNA: An in silico single molecule approach. *Biopolymers* **2007**, *85*, 169.
- (5) Trylska, J.; Tozzini, V.; McCammon, J. A. Exploring global motions and correlations in the ribosome. *Biophys. J.* **2005**, *89*, 1455.
- (6) Zagrovic, B.; Pande, V. Solvent viscosity dependence of the folding rate of a small protein: Distributed computing study. *J. Comput. Chem.* **2003**, *24*, 1432.
- (7) Russell, R.; Millett, I. S.; Tate, M. W.; Kwok, L. W.; Nakatani, B.; Gruner, S. M.; Mochrie, S. G. J.; Pande, V.; Doniach, S.; Herschlag, D.; Pollack, L. Rapid compaction during RNA folding. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 4266.
- (8) Zagrovic, B.; Sorin, E. J.; Pande, V. Beta-hairpin folding simulations in atomistic detail using an implicit solvent model. *J. Mol. Biol.* **2001**, *313*, 151.
- (9) Duan, Y.; Kollman, P. A. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **1998**, *282*, 740.
- (10) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179.
- (11) 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.
- (12) Scott, W. R. P.; Hunenberger, P. H.; Tironi, I. G.; Mark, A. E.; Billeter, S. R.; Fennel, J.; Torda, A. E.; Huber, T.; Kruger, P.; van Gunsteren, W. F. The GROMOS biomolecular simulation program package. *J. Phys. Chem. A* **1999**, *103*, 3596.
- (13) Sponer, J.; Riley, K. E.; Hobza, P. Nature and magnitude of aromatic stacking of nucleic acid bases. *Phys. Chem. Chem. Phys.* **2008**, *10*, 2595.
- (14) Gautheret, D. F.; Konings, D.; Gutell, R. R. A major family of motifs involving G•A mismatches in ribosomal RNA. *J. Mol. Biol.* **1994**, *242*, 1.
- (15) Gutell, R. R.; Gray, M. W.; Schnare, M. N. A compilation of large subunit (23S- and 23S-like) ribosomal-RNA structures: 1993. *Nucleic Acids Res.* **1993**, *21*, 3055.
- (16) Gutell, R. R.; Weiser, B.; Woese, C. R.; Noller, H. F. Comparative anatomy of 16S-like ribosomal-RNA. *Prog. Nucleic Acid Res. Mol. Biol.* **1985**, *32*, 155.
- (17) Cannone, J. J.; Subramanian, S.; Schnare, M. N.; Collett, J. R.; D'Souza, L. M.; Du, Y. S.; Feng, B.; Lin, N.; Madabusi, L. V.; Muller, K. M.; Pande, N.; Shang, Z. D.; Yu, N.; Gutell, R. R. The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics* **2002**, *3*, 2.
- (18) Pley, H. W.; Flaherty, K. M.; McKay, D. B. 3-Dimensional structure of a hammerhead ribozyme. *Nature* **1994**, *372*, 68.
- (19) Pley, H. W.; Flaherty, K. M.; McKay, D. B. Model for an RNA tertiary interaction from the structure of an intermolecular complex between a GAAA tetraloop and an RNA helix. *Nature* **1994**, *372*, 111.
- (20) Biou, V.; Yaremchuk, A.; Tukalo, M.; Cusack, S. The 2.9 Angstrom crystal structure of *T. thermophilus* Seryl-tRNA synthetase complexed with tRNA(Ser). *Science* **1994**, *263*, 1404.
- (21) Murphy, F. L.; Cech, T. R. GAAA tetraloop and conserved bulge stabilize tertiary structure of a Group-I Intron domain. *J. Mol. Biol.* **1994**, *236*, 49.

- (22) Michel, F.; Westhof, E. Modeling of the 3-Dimensional architecture of Group-I Catalytic Introns based on comparative sequence-analysis. *J. Mol. Biol.* **1990**, *216*, 585.
- (23) Zwieb, C. Recognition of a tetranucleotide loop of signal recognition particle RNA by protein-Srp19. *J. Biol. Chem.* **1992**, *267*, 15650.
- (24) SantaLucia, J.; Kierzek, R.; Turner, D. H. Effects of GA mismatches on the structure and thermodynamics of RNA internal loops. *Biochemistry* **1990**, *29*, 8813.
- (25) SantaLucia, J., Jr.; Turner, D. H. Structure of r(GGCGAGCC)₂ in solution from NMR and restrained molecular dynamics. *Biochemistry* **1993**, *32*, 12612.
- (26) Wu, M.; Turner, D. H. Solution structure of r(GCGGACGC)₂ by two-dimensional NMR and the iterative relaxation matrix approach. *Biochemistry* **1996**, *35*, 9677.
- (27) Villescás-Díaz, G.; Zacharias, M. Sequence context dependence of tandem Guanine-Adenine mismatch conformations in RNA: A continuum solvent analysis. *Biophys. J.* **2003**, *85*, 416.
- (28) Yildirim, I.; Turner, D. H. RNA challenges for computational chemists. *Biochemistry* **2005**, *44*, 13225.
- (29) Kollman, P. A. Free energy calculations: Applications to chemical and biochemical phenomena. *Chem. Rev.* **1993**, *93*, 2395.
- (30) Chen, G.; Kierzek, R.; Yildirim, I.; Krugh, T. R.; Turner, D. H.; Kennedy, S. D. Stacking effects on local structure in RNA: Changes in the structure of tandem GA pairs when flanking GC pairs are replaced by isoG-isoC pairs. *J. Phys. Chem. B* **2007**, *111*, 6718.
- (31) Hobza, P.; Sponer, J. Structure, energetics, and dynamics of the nucleic acid base pairs: Nonempirical ab initio calculations. *Chem. Rev.* **1999**, *99*, 3247.
- (32) Sponer, J.; Mokdad, A.; Sponer, J. E.; Spackova, N.; Leszczynski, J.; Leontis, N. B. Unique tertiary and neighbor interactions determine conservation patterns of cis Watson-Crick A/G base-pairs. *J. Mol. Biol.* **2003**, *330*, 967.
- (33) Sponer, J.; Florian, J.; Hobza, P.; Leszczynski, J. Nonplanar DNA base pairs. *J. Biomol. Struct. Dyn.* **1996**, *13*, 827.
- (34) Wang, J. M.; Cieplak, P.; Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules. *J. Comput. Chem.* **2000**, *21*, 1049.
- (35) Case, D. A.; Darden, T. A.; Cheatham, T. E. I.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Matthews, D. H.; Schafmeister, C.; Ross, W. S.; Kollman, P. A. *AMBER 9*, University of California: San Francisco, CA; 2006.
- (36) Simonson, T.; Carlsson, J.; Case, D. A. Proton binding to proteins: pK(a) calculations with explicit and implicit solvent models. *J. Am. Chem. Soc.* **2004**, *126*, 4167.
- (37) Shirts, M. R.; Pitera, J. W.; Swope, W. C.; Pande, V. S. Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins. *J. Chem. Phys.* **2003**, *119*, 5740.
- (38) Simonson, T. Free-Energy of particle insertion - an exact analysis of the origin singularity for simple liquids. *Mol. Phys.* **1993**, *80*, 441.
- (39) Wolfram Research, Inc. *Mathematica Edition*, Version 5.2; Wolfram Research, Inc.: Champaign, IL, 2005.
- (40) Pitera, J. W.; Van Gunsteren, W. F. A comparison of non-bonded scaling approaches for free energy calculations. *Mol. Simul.* **2002**, *28*, 45.
- (41) Zacharias, M.; Straatsma, T. P.; McCammon, J. A. Separation-shifted scaling, a new scaling method for Lennard-Jones interactions in thermodynamic integration. *J. Chem. Phys.* **1994**, *100*, 9025.
- (42) 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.
- (43) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Kollman, P. A. Application of RESP charges to calculate conformational energies, hydrogen-bond energies, and free-energies of solvation. *J. Am. Chem. Soc.* **1993**, *115*, 9620.
- (44) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges - the Resp Model. *J. Phys. Chem.* **1993**, *97*, 10269.
- (45) 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.
- (46) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926.
- (47) Ryjacek, F.; Kubar, T.; Hobza, P. New parameterization of the Cornell et al. empirical force field covering amino group nonplanarity in nucleic acid bases. *J. Comput. Chem.* **2003**, *24*, 1891.
- (48) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical-Integration of cartesian equations of motion of a system with constraints: Molecular-Dynamics of N-Alkanes. *J. Comput. Phys.* **1977**, *23*, 327.
- (49) Jeffrey, G. A. *An Introduction to Hydrogen Bonding*; Oxford University Press: New York, 1997; pp 11–32.
- (50) Chen, X.; Kierzek, R.; Turner, D. H. Stability and structure of RNA duplexes containing isoguanosine and isocytidine. *J. Am. Chem. Soc.* **2001**, *123*, 1267.
- (51) Leontis, N. B.; Westhof, E. Geometric nomenclature and classification of RNA base pairs. *RNA* **2001**, *7*, 499.
- (52) Grossfield, A.; Ren, P. Y.; Ponder, J. W. Ion solvation thermodynamics from simulation with a polarizable force field. *J. Am. Chem. Soc.* **2003**, *125*, 15671.

- (53) Ren, P. Y.; Ponder, J. W. Polarizable atomic multipole water model for molecular mechanics simulation. *J. Phys. Chem. B* **2003**, *107*, 5933.
- (54) Babin, V.; Baucom, J.; Darden, T. A.; Sagui, C. Molecular dynamics simulations of DNA with polarizable force fields: Convergence of an ideal B-DNA structure to the crystallographic structure. *J. Phys. Chem. B* **2006**, *110*, 11571.
- (55) Gresh, N.; Sponer, J. E.; Spackova, N.; Leszczynski, J.; Sponer, J. Theoretical study of binding of hydrated Zn(II) and Mg(II) cations to 5'-guanosine monophosphate. Toward polarizable molecular mechanics for DNA and RNA. *J. Phys. Chem. B* **2003**, *107*, 8669.
- (56) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **1996**, *14*, 33.

CT800540C

A Density Functional Theory Study of Ground and Low-Lying Excited Electronic States in Defective Graphenes

Hiroto Tachikawa,* Yoshinori Nagoya, and Hiroshi Kawabata

Division of Materials Chemistry, Graduate School of Engineering, Hokkaido University, Sapporo 060-8628, Japan

Received March 30, 2009

Abstract: Electronic states of graphenes, whose carbon atoms are terminated by hydrogen atoms (hydrogenated graphene, denoted H-graphene) and defective graphene (one carbon atom was removed from H-graphene, denoted D-graphene) have been investigated by density functional theory. The sizes of graphenes examined in the present study were $n = 7, 14, 19, 29, 37, 44,$ and 52 ; where n is the number of benzene rings in the graphene. The excitation energies of H-graphenes were gradually decreased as a function of the number of rings. In D-graphene, new energy levels for the first and second excited states appeared as low-lying excited states. It was found that the formation of defect sites in graphene produces large decreases in the excitation energies for third and higher excited states. The highest occupied molecular orbital and lowest unoccupied molecular orbital (LUMO) in H-graphene were widely delocalized over the graphene surface. On the other hand, LUMO in D-graphene was localized only in the defect sites. The effects of vacancy defects on both the ground and excited electronic states of graphene were discussed on the basis of theoretical results.

1. Introduction

Graphene is a single layer composed of a carbon backbone with a benzene ring structure. An ideal graphene is a semimetal or zero-energy gap semiconductor. It is expected that the graphene can be applied to an electronic device and circuit because it shows electronic conductivity that is higher than that of three-dimensional silicon.^{1–3} Experimental results from transport measurements show that graphene has a remarkably high electron mobility at room temperature, with reported values in excess of $1.5 \times 10^4 - 2.0 \times 10^5$ cm²/(V s).^{4–6} In addition, the symmetry of the experimentally measured conductance indicates that the mobilities for holes and electrons are nearly the same. The mobility is nearly independent of temperature between 10 and 100 K.^{7–9}

Existence of defect vacancy in graphene and carbon nanotube has been pointed out by several groups.^{10–22} Acid treatment is known to create defects in carbon materials.^{13–15} The C–C bond cleavage is caused by electrophilic attack.

Coleman et al. artificially made the defect in graphene by acid treatment.¹⁹ They treated the graphene surface by using nonoxidizing acid (HCl) and found the carbon vacancy defect with an in-plane orbital. They also carried out band calculations for two vacancy sites and showed that the band gap in defect site is significantly small.

The electronic states of defect in the graphene have been investigated from a theoretical point of view. Yazyev and Helm investigated the magnetism in graphene induced by single carbon atom defects.^{11,21} They considered two types of defects, i.e., the hydrogen chemisorption defect and the vacancy defect. The magnetism due to the defect-induced extended states has been found. Using first-principles calculation, Duplock et al. investigated the adsorption of atomic hydrogen on the defect of graphene. They found that a new energy gap was opened due to one of the electronic density of states.²² Thus, the electronic states of the defect of graphene at the ground state were well understood theoretically. However, electronic states at both ground and excited states of defective graphene are scarcely known.

* Corresponding author. Fax: 81-11-706-7897. E-mail: hiroto@eng.hokudai.ac.jp.

In the present study, ground and low-lying excited states of normal and defective (one carbon atom is removed from the normal graphene) graphene have been investigated by means of density functional theory (DFT) method^{23–26} to elucidate the effect of the defect on the electronic states of graphene. In particular, we focus our attention on the excited states of vacancy defect in graphenes.

The article is organized as follows. In the Computational Details Section, we present the structural models of nano-graphenes, the method used in the geometry optimization, and the electronic structure calculations of graphenes. Results are described in the Results Section. Finally, the Discussion Section gives the model of the electronic excitation of defective graphene, and our conclusions are presented in the Conclusion Section.

2. Computational Details

In the present study, normal graphene (denoted by H-graphene) and defective graphene (D-graphene) were examined to elucidate the effect of vacancy defect on the electronic states of graphene. The edges of graphenes were terminated by hydrogen atoms. The defective graphene was made by removing one carbon atom from the graphene. Namely, D-graphene has a vacancy defect in the central region of the graphene. Seven graphenes, $n = 7, 14, 19, 29, 37, 44,$ and 52 , were examined as models of graphene; where n is the number of benzene rings in graphene. The structures of all graphenes are given in the Supporting Information. The closed-shell singlet states were examined in the present work.

The structures of the graphenes were optimized at the B3LYP/6-31G(d) level of theory. The excitation energies of H-graphene and D-graphene were calculated at the time-dependent (TD)-DFT, B3LYP/6-31G(d) level. In addition, the PW91PW91/3-21G(d) method was used for comparison. To check the basis set and method dependency, the electronic states were calculated by means of the PW91PW91 level for $n = 7–52$ and the B3LYP/6-311G(d) level for $n = 7–19$. Similar results were obtained, suggesting that one can discuss the qualitative feature of electronic states of graphene using the B3LYP/6-31G(d) level. All DFT calculations were carried out using the Gaussian 03 program package.²⁷

To check the validity of the B3LYP/6-31G(d) surface, the B3LYP/6-311G(d) calculations were carried out for the defective graphene with $n = 37$. The special distributions of molecular orbitals (MOs) around the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) obtained by the B3LYP/6-311G(d) level were found to be significantly close to those of the B3LYP/6-31G(d) calculations (see Supporting Information). Therefore, the B3LYP/6-31G(d) level of theory is judged to be effective to obtain qualitative features of the present system.

3. Results

3.1. Electronic States of Hydrogen Terminated Graphene (H-graphene). The structures of the hydrogen terminated graphene (denoted H-graphene) used in the present study were fully optimized at the B3LYP/6-31G(d) level. For

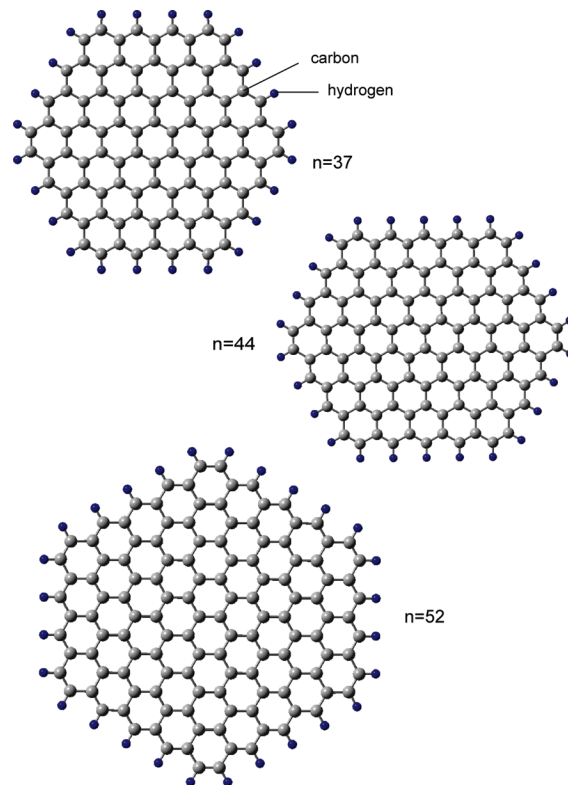


Figure 1. Optimized structure of graphenes (H-graphene) obtained by at the B3LYP/6-31G(d) level. Only graphenes with $n = 37, 44,$ and 52 are given. Edge region of graphene is terminated by hydrogen atoms.

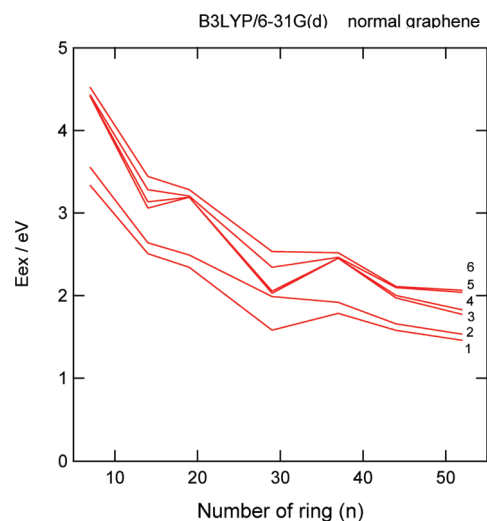


Figure 2. Excitation energies of H-graphenes plotted as a function of the size of graphene (n). The energies are calculated at the TD-B3LYP/6-31G(d) level. Six electronic excited states are solved.

example, three optimized structures of $n = 37, 44,$ and 52 are given in Figure 1.

The excitation energies of H-graphenes were calculated as a function of size (n) at the TD-B3LYP/6-31G(d) level. The results are given in Figure 2. The excitation energies of $n = 7$ (smallest size) were calculated to be 3.34 (1st), 3.56 (2nd), and 4.42 eV (3rd excitation energy). It was found that the excitation energy decreases with increasing n ; for example, the first excitation energies for $n = 14, 19, 37,$

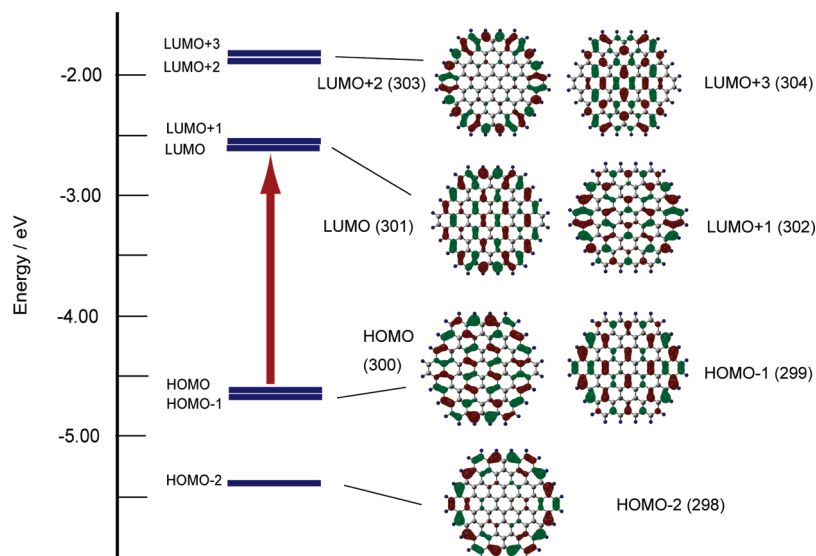


Figure 3. MO energies of H-graphenes calculated at the B3LYP/6-31G(d) level. Isosurface indicates special distributions of the molecular orbitals of H-graphene. Arrow indicates main configuration of the first electronic excitations. Orbital numbers are given in parentheses.

Table 1. Excitation Energies and Configuration State Functions of H-graphene ($n = 37$) Calculated at the TD-DFT(B3LYP)/6-31G(d) Level

state	CSF and CI vectors	E_{ex}, eV
S_1	$0.515\phi(\text{HOMO-1} \rightarrow \text{LUMO}) - 0.515\phi(\text{HOMO} \rightarrow \text{LUMO+1})$	1.67
S_2	$0.492\phi(\text{HOMO-1} \rightarrow \text{LUMO+1}) - 0.492\phi(\text{HOMO} \rightarrow \text{LUMO})$	1.81
S_3	$0.331\phi(\text{HOMO} \rightarrow \text{LUMO+4}) - 0.331\phi(\text{HOMO-1} \rightarrow \text{LUMO+3})$	2.34
S_4	$0.371\phi(\text{HOMO} \rightarrow \text{LUMO+3}) + 0.331\phi(\text{HOMO-1} \rightarrow \text{LUMO+4})$	2.37

and 52 were calculated to be 2.51, 2.34, 1.69, and 1.46 eV, respectively. The shape of energy curve indicates that the energy gap decreases gradually with increasing n , and it is almost saturated around $n = 29$ –52.

The MOs of H-graphene ($n = 37$, for example) are illustrated in Figure 3 together with the orbital energies around HOMO and LUMO calculated at the B3LYP/6-31G(d) level. Both HOMO and LUMO are doubly degenerated in energy. The orbitals from HOMO-1 and LUMO+1 are widely delocalized over the graphene surface. On the other hand, those of HOMO-2 and LUMO+2 are localized in edge region of graphene surface. In the case of LUMO+3, the orbital is delocalized again on the surface.

To elucidate the electronic states of graphene at the ground and low-lying excited states, MOs and weights of configuration state functions (CSFs) are analyzed in detail. The excitation energies and CSFs are given in Table 1. The main configurations for the first and second excitations are $\phi(\text{HOMO-1} \rightarrow \text{LUMO})$ and $\phi(\text{HOMO} \rightarrow \text{LUMO+1})$, while the coefficients of CSFs are 0.515 and -0.515 , respectively. Here, $\phi(\text{HOMO} \rightarrow \text{LUMO})$ means a CSF where one electron is excited from HOMO to LUMO.

It should be noted that HOMO and HOMO-1 are doubly degenerated, while LUMO and LUMO+1 are also doubly degenerated. These results indicate that the low-lying excited states of graphene are composed of the HOMO–LUMO excitations.

3.2. Electronic States of D-graphene. To elucidate the effects of the defect on the electronic states of graphene,

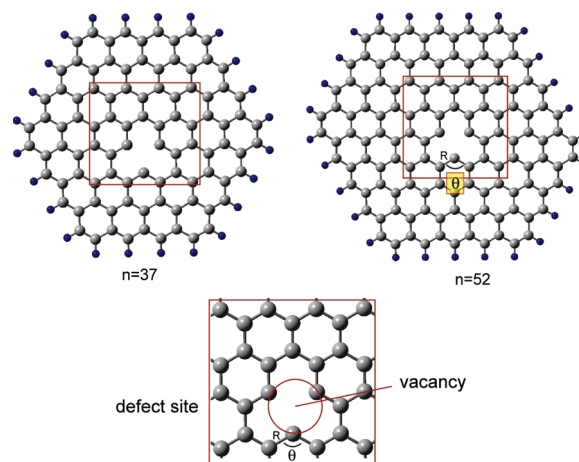


Figure 4. Optimized structure of D-graphene obtained by at the B3LYP/6-31G(d) level. Only D-graphenes with $n = 37$, 44, and 52 are given. Edge region of graphene is terminated by hydrogen atoms.

a defect was artificially made on the graphene. The defect is made by removing one carbon atom located around the center of H-graphene. The geometry of each D-graphene was optimized at the B3LYP/6-31G(d) level. The structures of the D-graphenes for $n = 37$ and 52 are illustrated in Figure 4.

The stabilization energy obtained by the geometry optimization was calculated to be 14.0 kcal/mol ($n = 37$). The C–C bond length ($=R$) and C–C–C angle ($=\theta$) in the defect site were changed from 1.425 Å and 120.1° to 1.393 Å and 123.8° , respectively. The bond lengths were slightly shortened and the angle is wider by the deformation.

The excitation energies of defective graphenes ($n = 7, 14, 19, 37, 44$, and 52) are plotted in Figure 5 as a function of n . The energy level for $n = 29$ was not given because the TD calculation was not converged in case of $n = 29$. The excitation energy decreased with increasing n as well as that of H-graphene. It was found that the excitation energies of

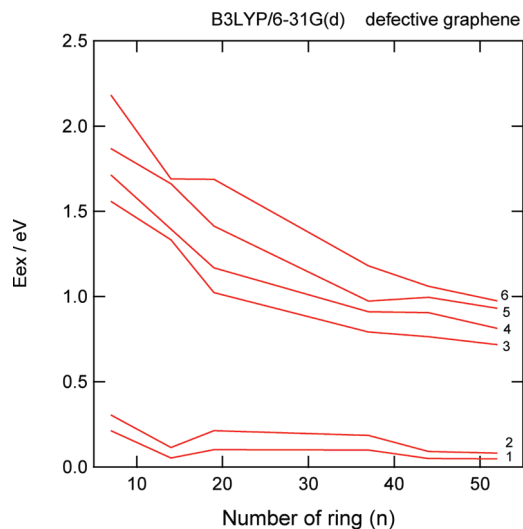


Figure 5. Excitation energies of D-graphenes plotted as a function of size of graphene (n). The energies are calculated at the TD-B3LYP/6-31G(d) level. Six electronic excited states are solved.

the defective graphene at the first and second excitations were significantly lower than those of H-graphene (below 0.15 eV). Also, it is shown that the first and second excitation energies were not dependent on size (n). In contrast, third and higher excitation energies in the defective graphenes were strongly dependent on n , and these energy levels were also lower than those of H-graphene; for example, the third excitation energy for the defective graphene with $n = 52$ is 0.72 eV, while that of H-graphene with $n = 52$ is 1.77 eV.

Thus, it was found that new energy levels (first and second excited states) appear by the formation of a vacancy defect in graphene. The energy levels for the first and second excitations of D-graphenes are significantly lower than those of H-graphenes. These excitations are not dependent on the size of graphene. On the other hand, the third and higher

Table 2. Excitation Energies and Configuration State Functions of D-graphene ($n = 37$) Calculated at the TD-DFT(B3LYP)/6-31G(d) Level

state	CSF and CI vectors	E_{ex} , eV
S_1	$0.614\phi(\text{HOMO} \rightarrow \text{LUMO}) - 0.126\phi(\text{HOMO-2} \rightarrow \text{LUMO})$	0.405
S_2	$0.631\phi(\text{HOMO} \rightarrow \text{LUMO+1}) - 0.156\phi(\text{HOMO-2} \rightarrow \text{LUMO+1})$	0.468
S_3	$0.694\phi(\text{HOMO-1} \rightarrow \text{LUMO})$	1.035
S_4	$0.696\phi(\text{HOMO-1} \rightarrow \text{LUMO+1})$	1.067

excitation energies are strongly dependent on cluster size (n). The lower energy levels obtained in the D-graphene are much different from that of H-graphenes.

To elucidate the specific electronic feature in D-graphene, molecular orbitals (MOs) and weights of configuration state functions (CSFs) at the excited states are analyzed in detail. The special distributions of MOs and orbital energies around HOMO and LUMO are given in Figure 6. It was found that HOMO, HOMO-1, and HOMO-2 are widely delocalized on graphene surface. On the other hand, LUMO and LUMO+1 are localized in the defect site. This feature is much different from the H-graphene in which both HOMO and LUMO are delocalized on the graphene surface. These results indicate that formation of vacancy defect causes a large change of electronic states of graphene. Especially, it can be expected that the excited and anionic states of graphene are strongly affected by the defect.

The coefficients of CSFs in the D-graphene are summarized in Table 2. The first excited state was mainly composed of two CSFs, $\phi(\text{HOMO} \rightarrow \text{LUMO+1})$ and $\phi(\text{HOMO-2} \rightarrow \text{LUMO})$, while the coefficients of CSFs for the former and the latter were calculated to be 0.914 and -0.126 , respectively. The second excited state is mainly composed of $\phi(\text{HOMO} \rightarrow \text{LUMO+1})$ with the coefficient of 0.631. The first and second excitations are almost degenerated in from energy each other. Therefore, the first excitation band ($S_0 \rightarrow S_1$ and $S_0 \rightarrow S_2$) can be assigned to a charge transfer (CT) band from the bulk surface to the defect

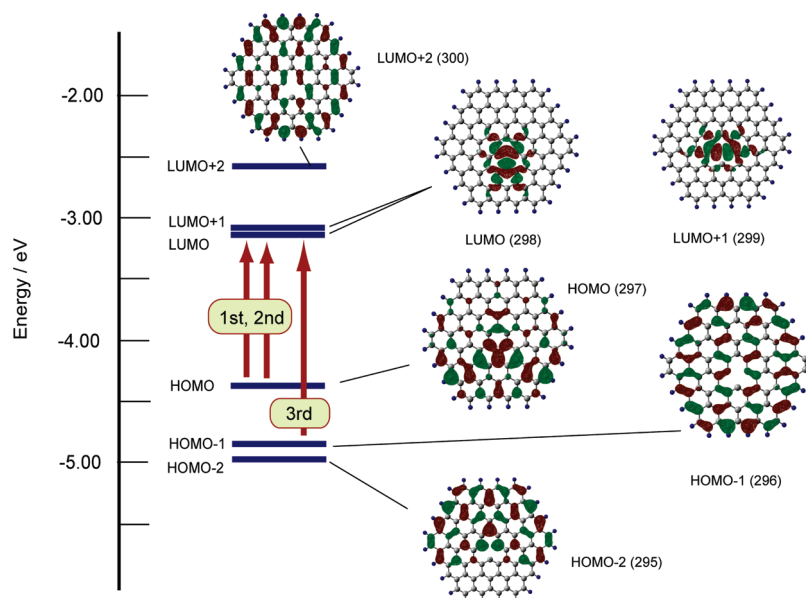


Figure 6. Molecular orbital energies of D-graphenes calculated at the B3LYP/6-31G(d) level. Isosurface indicates the special distributions of molecular orbitals of D-graphene. Arrows indicate main configurations of first, second, and third electronic excitations. Orbital numbers are given in parentheses.

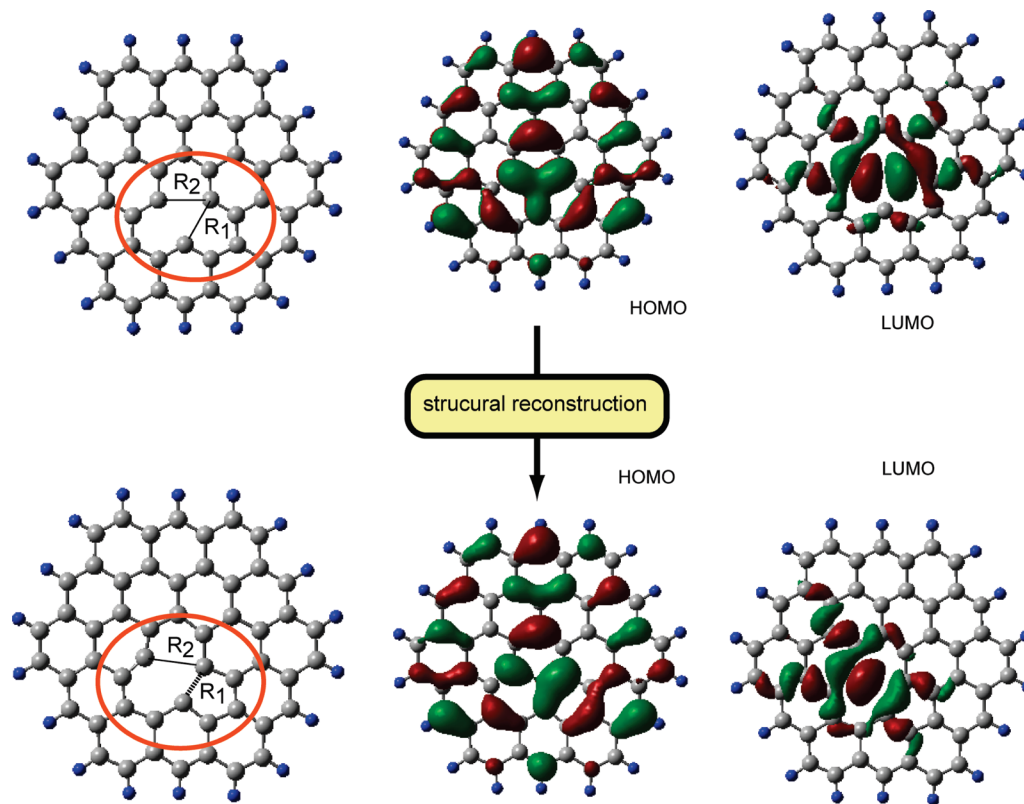


Figure 7. Special distributions of HOMO and LUMO of D-graphenes before and after structural reconstruction.

site. The third and fourth excitation bands are also attributed to one of the CT bands.

3.3. Effects of Structural Reconstruction of D-graphene on the Electronic States. It is known that a vacancy defect undergoes reconstruction by thermal activation. In this section, the effects of reconstruction on the excitation energies are discussed. The optimized structures of D-graphene ($n = 19$) before and after structural reconstruction are illustrated in Figure 7. The large vacancy with three benzene rings is formed before structural reconstruction. The bond distances between carbon atoms are $R_1 = 2.433 \text{ \AA}$ and $R_2 = 2.463 \text{ \AA}$. After the reconstruction, the bond distances are changed to $R_1 = 1.866 \text{ \AA}$ and $R_2 = 2.694 \text{ \AA}$, indicating that one of the C–C bonds is significantly shortened and a new C–C bond is formed by the reconstruction. The stabilization energy is 14.4 kcal/mol at the B3LYP/6-31G(d) level. However, distributions of HOMO and LUMO show that the shapes of orbitals are not strongly affected by the reconstruction. The excitation energies of D-graphene are calculated to be 0.435 (1st) and 0.566 eV (2nd) before the reconstruction and 0.290 (1st) and 0.470 eV (2nd) after the reconstruction, indicating that the reconstruction affects the excitation energies. However, the assignment of the excitation is not changed.

4. Discussion

4.1. Model of Electronic Excitation in D-graphene. In the present study, the electronic states of D-graphene have been investigated by means of the DFT method. A single vacancy defect was examined and compared with pure

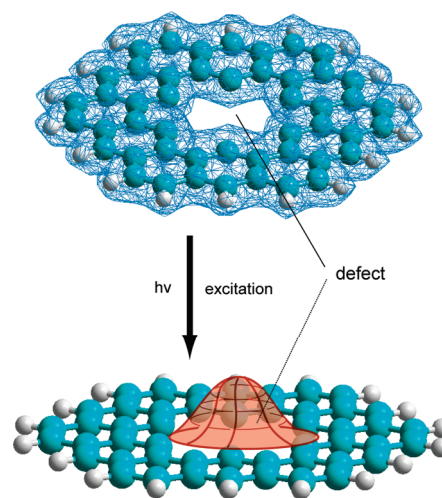


Figure 8. Schematic illustration of a model of electronic excitations of D-graphene. The electron is concentrated in the defect site by the electronic excitation.

graphene. It was found that low-lying excited states newly appear by the formation of defect. These excited states are composed of electron transfer bands from the bulk surface to the defect site.

On the basis of the present results, we propose a model for the electronic excitations of D-graphene, which is illustrated schematically in Figure 8. The D-graphene has new energy bands at lower energy levels below 0.15 eV. HOMO of the D-graphene is distributed widely on the graphene surface as shown in Figure 8 (upper). After the electronic excitation of D-graphene, the electron is transferred from the bulk region to the defect site. Namely, it is found

that the D-graphene has a lowest excitation band corresponding to a charge transfer band from the bulk to defect site.

4.2. Comparison with Previous Studies. Recently, Coleman et al. investigated experimentally a defect vacancy in a graphene sheet.¹⁹ They treated the graphene sheet with an acid treatment (HCl) and made a vacancy defect in the graphene surface. It was suggested that defect vacancy has a metallic conductivity. The present calculation strongly supports this feature; namely, the band gap of graphene is significantly decreased by the formation of a vacancy defect. The first excitation energy is changed from 1.46 to 0.05 eV by the formation of a vacancy defect ($n = 52$) at the B3LYP/6-31G(d) level.

In the present study, the closed-shell singlet states were discussed only as an electronic structure. However, it has been pointed out that the spin polarization is important in open-shell singlet states of nanosized graphenes.²⁸ Jiang and Dai investigated that the electronic ground states of acenes with different numbers of fused benzene rings (up to 40 rings) have been studied with first principles spin-polarized DFT. The ground states of higher acenes (>7 rings) are found to be antiferromagnetic (i.e., open-shell singlet). To elucidate the electronic structures of graphenes in details, the calculations of open-shell singlet excited states are needed in the near future. For comparison, the spin density of graphene ($n = 37$) at the triplet states was calculated. The distribution of spin density is illustrated in Supporting Information (Figure S7). The unpaired electrons were mainly localized in the defect site, but the tail of the distribution is widely delocalized over the graphene surface. However, the energy of triplet state was close to that of the singlet ground state (the T_1 state was only 0.3 kcal/mol higher in energy than the S_0 state). Therefore, the investigation of electronic states at the triplet states will be needed in the near future to elucidate the detailed electronic structure of D-graphene.

It has been pointed out that the TD-DFT method without Hartree–Fock (HF) exchange misses completely the excitonic effects. The magnitude of the error would be dependent on the system. There are cases where the error can be quite significant.²⁹ In the case of a large system, the effects may be large.³⁰ To validate the TD-DFT(B3LYP) calculations in the present system, time-dependent HF (TDHF) calculations^{31,32} have been done for comparison. The results are given in the Supporting Information. The shapes of MOs around the HOMO are in good agreement with the TD-DFT calculations. The excitation energies of D-graphene calculated by TDHF were slightly larger than those of TDDFT. The assignment of electronic excitations is agreement with each other. Therefore, TD-DFT(B3LYP) calculation would give a reasonable electronic structure of the D-graphene in a qualitative feature.

5. Conclusion

In the present study, the electronic states of normal and defective graphenes (denoted by H- and D-graphenes, respectively) have been investigated by density functional theory (DFT) calculations. The sizes of graphenes examined in the present study were $n = 7, 14, 19, 29, 37, 44,$ and 52 ; where n is the number of benzene rings in the graphene.

The excitation energies of H-graphenes were gradually decreased as a function of the number of rings. In D-graphene, new energy levels for the first and second excited states appeared as low-lying excited states. The formation of defect sites in graphene produces large decreases in the excitation energies for third and higher excited states. The first excitation band of D-graphene is assigned as a charge transfer band from the bulk to the defect site.

Acknowledgment. H.T. is indebted to the Computer Center at the Institute for Molecular Science (IMS) for the use of the computing facilities. He also acknowledges partial support from the Iketani Foundation and a grant-in-aid for Scientific Research (C) from the Japan Society for the Promotion of Science (JSPS). This work was partially supported by the KURATA foundation.

Supporting Information Available: Illustrations of optimized structures of D-graphene ($n = 7, 14, 19, 29, 37, 44,$ and 52), and excitation energies of H- and D-graphenes were calculated at several levels of theory. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Geim, A. K.; Novoselov, K. S. The rise of graphene. *Nat. Mater.* **2007**, *6*, 183–191.
- (2) Novoselov, K. S. *Nat. Mater.* **2007**, *6*, 720–721.
- (3) Yang, L.; Cohen, M. L.; Louie, S. G. *Phys. Rev. Lett.* **2008**, *101*, 186401.
- (4) Morozov, S. V.; Novoselov, K. S.; Katsnelson, M. I.; Schedin, F.; Elias, D. C.; Jaszczak, J. A.; Geim, A. K. *Phys. Rev. Lett.* **2008**, *100*, 016602.
- (5) Bolotin, K. I.; Sikes, K. J.; Hone, J.; Stormer, H. L.; Kim, P. *Phys. Rev. Lett.* **2008**, *101*, 096802.
- (6) Chen, J.-H.; Jang, C.; Xiao, S.; Ishigami, M.; Fuhrer, M. S. *Nature Nanotechnol.* **2008**, *3*, 206.
- (7) Novoselov, K. S.; Geim, A. K.; Morozov, S. V.; Jiang, D.; Katsnelson, M. I.; Grigorieva, I. V.; Dubonosand, S. V.; Firsov, A. A. *Nature* **2005**, *438*, 197.
- (8) Wehling, T. O.; Novoselov, K. S.; Morozov, S. V.; Vdovin, E. E.; Katsnelson, M. I.; Geim, A. K.; Lichtenstein, A. I. *Nano Lett.* **2008**, *8*, 173–177.
- (9) Hwang, E. H.; Das Sarma, S. *Phys. Rev.* **2009**, *79*, 165404.
- (10) Khare, R.; Mielke, S. L.; Paci, J. T.; Zhang, S.; Ballarini, R.; Schatz, G. C.; Belytschko, T. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2007**, *75*, 075412.
- (11) Yazyev, O. V.; Helm, L. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2007**, *75*, 125408.
- (12) Rocha, A. R.; Padilha, J. E.; Fazzio, A.; da Silva, A. J. R. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2008**, *77*, 153406.
- (13) Li, T. C.; Lu, S.-P. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2007**, *77*, 085408.
- (14) Pimenta, M. A.; Dresselhaus, G.; Dresselhaus, M. S.; Cancado, L. G.; Jorio, A.; Saito, R. *Phys. Chem. Chem. Phys.* **2007**, *9*, 1276–1291.
- (15) Kuemmeth, F.; Ilani, S.; Ralph, D. C.; McEuen, P. L. *Nature* **2008**, *452*, 448–452.

- (16) Balasubramanian, K.; Burghard, M. *Small* **2005**, *1*, 180–192.
- (17) Itkis, M. E.; Perea, D. E.; Jung, R.; Niyogi, S.; Haddon, R. C. *J. Am. Chem. Soc.* **2005**, *127*, 3439–3448.
- (18) Strano, M. S.; Huffman, C. B.; Moore, V. C.; O’Connell, M. J.; Haroz, E. H.; Hubbard, J.; Miller, M.; Rialon, K.; Kittrell, C.; Ramesh, S.; Hauge, R. H.; Smalley, R. E. *J. Phys. Chem. B* **2003**, *107*, 69796985.
- (19) Coleman, V. A.; Knut, R.; Karis, O.; Grennberg, H.; Jansson, U.; Quinlan, R.; Holloway, B. C.; Sanyal, B.; Eriksson, O. *J. Phys. D: Appl. Phys.* **2008**, *41*, 062001.
- (20) Boukhalov, D. W.; Katsnelson, M. I. *Nano Lett* **2008**, *8*, 4373–4379.
- (21) Yazyev, O. V.; Helm, L. *J. Phys. Conf. Ser.* **2007**, *61*, 1294–1298.
- (22) Duplock, E. J.; Scheffler, M.; Lindan, P. J. D. *Phys. Rev. Lett.* **2004**, *92*, 225502.
- (23) Tachikawa, H.; Shimizu, A. *J. Phys. Chem. B* **2005**, *109*, 13255–13262.
- (24) Tachikawa, H.; Shimizu, A. *J. Phys. Chem. B* **2006**, *110*, 20445–20450.
- (25) Tachikawa, H. *J. Phys. Chem. C* **2007**, *111*, 13087–13091.
- (26) Tachikawa, H. *J. Phys. Chem. C* **2008**, *112*, 10193–10199.
- (27) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, Revision B.04; Gaussian, Inc.: Pittsburgh PA, 2003.
- (28) Jiang, D.; Dai, S. *J. Phys. Chem. A* **2008**, *112*, 332–335.
- (29) Yang, L.; Cohen, M. L.; Louie, S. G. *Nano Lett.* **2007**, *7*, 3112–3115.
- (30) Izmaylov, A. F.; Scuseria, G. E. *J. Chem. Phys.* **2008**, *129*, 034101.
- (31) Hieringer, W.; Gorling, A. *Chem. Phys. Lett.* **2006**, *426*, 234–236.
- (32) Hieringer, W.; Gorling, A. *Chem. Phys. Lett.* **2006**, *419*, 557–562.

CT900151S

A Quantum Chemical Interpretation of Compressibility in Solids

J. Contreras-García,[†] P. Mori-Sánchez,[‡] B. Silvi,[§] and J. M. Recio^{*,†}

*MALTA-Consolider Team and Departamento de Química Física y Analítica
Universidad de Oviedo, E-33006 Oviedo, Spain; Department of Chemistry, Duke
University, Durham, North Carolina 27708; Laboratoire de Chimie Théorique,
Université Pierre et Marie Curie, F-75252 Paris, France*

Received May 5, 2009

Abstract: The ability of the electron localization function to perform a partition of the unit cell volume of crystalline solids into well-defined, disjoint, and space-filling regions enables us to decompose the bulk compressibility into local contributions with a full chemical meaning. This partition has been applied to a set of prototype crystals of the chemical elements of the first three periods of the periodic table, and the equations of state for core, valence, bond, and lone electron pairs have been obtained. Solids are unequivocally classified into two groups according to their response to hydrostatic pressure. Those with sharing electrons (metals and covalent crystals) obey a simple relationship between the average valence electron density and the zero pressure bulk modulus. The stiffness of the closed-shell systems (molecular and ionic solids) is rationalized resorting to the Pauli principle. Overall, the results clearly correlate with chemical intuition: periodic trends are revealed, cores are almost incompressible and do not contribute appreciably to the macroscopic compressibility, and lone pair basins are rather easier to compress than bond basins.

1. Introduction

Since the beginning of the 20th century,^{1,2} many phenomenological equations and theoretical models have been proposed with the aim at explaining and predicting the compressibility of crystalline materials. Standard equations of state nowadays in use for solids have their roots in these pioneer works. State of the art developments (see ref 3 and references therein) highlighted the importance of finding correlations between the systematic behavior of the zero pressure bulk modulus (B_0) along the periodic table with the microscopic properties of chemical elements. Interestingly enough, Gilman³ invoked the core–valence separation to emphasize that the average valence electron density (the ratio between the valence electron population and the correspond-

ing volume) is a key parameter that shows an interesting correlation with B_0 in simple solid materials.

Though it is clear that the macroscopic compressibility of a crystal is ultimately determined by its chemical bonds and valence electrons, no systematic and quantitative account of this correlation in quantum chemical terms has been given so far. This is partially due to the need of an unambiguous partition of the unit cell volume into well-defined, disjoint, and space-filling regions. It is our point of view that such a partition, of the 3D space into core and valence shells with a subsequent division into regions associated with bond and lone electron pairs, can contribute to a global understanding of the compressibility of solids. In order to properly define the local compressibilities of the subregions, the volumes and charge populations of these basins have to be accessible. The topology of the electron localization function (ELF),⁴ which is derived from the conditional pair electron density and is related to its Laplacian, provides such a partition, and the integration of space and charge within the above basins allows us to assign volumes and electron populations to each

* Corresponding author. Fax: (+34)985 103125. E-mail: mateo@fluor.quimica.uniovi.es.

[†] Universidad de Oviedo.

[‡] Duke University.

[§] Université Pierre et Marie Curie.

of them. Thus, the equations of state and the average electron density for all these chemical regions will be at hand.

In this paper, our aim is to illustrate how the information contained in crystalline wave functions can be quantitatively translated to the macroscopic compressibility in terms of chemical (core, valence, bond, and lone pairs) contributions. Our target systems are the crystalline phases of most of the chemical elements located in the first three rows of the periodic table. We have computed volumes and charges at different pressures for all the regions or basins in which the ELF divides their respective unit cells. Systematic trends have been recovered, and general trends will be proposed. The validity of the empirical relationships based on the average electron density of pressure sensitive regions, as proposed by Gilman,³ is hereby tested by means of the quantitatively and theoretically grounded analysis of valence properties provided by the topological analysis of the ELF.

We organize the rest of the paper as follows. In Section 2, the ELF basic concepts and computational details are briefly summarized. Section 3 deals with the pictures that ELF provides to understand how pressure affects solids of different chemical bonding nature. Section 4 contains the results of our analysis on the response to pressure of the solids of most of the chemical elements with atomic numbers up to $Z = 20$. The main conclusions are gathered in Section 5.

2. Theoretical Background and Computational Details

2.1. Summary of ELF Concepts. Becke and Edgecombe⁴ introduced the ELF as a local measure of electron localization. Its core, χ , is formulated in terms of the conditional probability of finding one electron in the neighborhood of another one of the same spin, rescaled with respect to this probability in a uniform electron gas value of the same electron density. In order to have a direct relationship between the function and localization values of ELF are constrained between 0 and 1 by a *cosmetic* Lorentzian transformation $\eta = 1/(1 + \chi^2)$, so that when η tends to 1, parallel spins are highly improbable (and paired or lone electrons are, therefore, highly probable), whereas η tends to 0.5 in those regions where the electrons follow the homogeneous electron gas distribution.

A detailed presentation of the topological features of ELF is given elsewhere.⁵ We recall here that zero flux surfaces of the ELF gradient enclose 3D regions or basins that can be associated with electron shells, bonds, or lone pairs. The power of the ELF partition in the case of crystalline systems lies in the fact that these basins are finite, disjoint, and space filling, which means that the macroscopic volume is recovered when all of them are added together. Each of these basins contains a maximum or attractor of ELF, and between two of these attractors we find a first-order saddle point whose ELF value provides information about the interaction between the corresponding basins. If the ELF value at this point is very low, then electron delocalization between those basins is not expected. In practical terms, it is the same as to say that there will not be charge flow from one basin to the other basin when pressure is applied. The opposite is

expected if the ELF value is high. These first-order saddle points are called bond interaction points or *bips*. To facilitate the identification of these relevant relationships, it is interesting to track the ELF values along the path connecting two atoms of the unit cell. This plot highlights the approximate position of the attractors and the *bips* of the system revealing the atomic structure in shells, the presence of bonds and lone pairs, and how the different basins are interconnected. In the Section 2.2, ELF profiles for different bonding patterns will be analyzed.

2.2. Computational Parameters. In order to build the electron localization function for a given crystal structure, it is necessary to carry out all-electron calculations. We have followed first-principles methodologies based on the density functional theory as implemented in VASP and CRYSTAL98 packages.^{6,7} Since it is known to affect to a little extent the results provided by the topological analysis of ELF,⁸ the choice of functionals was mainly dictated by the proper computational convergence of the crystalline phase. As a general guideline, delocalized structures, such as metals, were calculated under the local density approximation,⁹ whereas the Perdew–Wang functional¹⁰ was preferred for the electronic description of ionic and molecular solids. Basis sets have been taken from literature sources,^{11,12} and standard computational parameters have been checked to ensure convergence in the calculations. We have computed with the VASP code the energy and the electronic structure at different unit cell volumes of the crystalline phases of most of the chemical elements with atomic numbers up to $Z = 20$. The energy–volume data for each of the compounds provide the necessary input to evaluate the corresponding equation of state (EOS) parameters: the equilibrium volume (V_0), the bulk modulus (B_0), and its first pressure derivative (B'_0) evaluated at zero pressure. Single point calculations at each of the full optimized structures calculated with VASP were later performed with CRYSTAL98 to obtain the total electron density in the unit cell from all-electron calculations.

An extension of our CRITIC code¹³ has been recently developed to carry out a thorough topological analysis of the ELF in crystals.^{14–17} Making use of automated and efficient recursive algorithms, it is able to provide an accurate analysis and characterization of critical points of the ELF topology in the unit cell. Basin volumes (v) and charges (q) are calculated by integrating the appropriate density operators and recovering the unit cell value. A thorough description of the new abilities of the CRITIC code in the studies of the ELF topology in crystals is given elsewhere.¹⁷

For each of the crystalline phases under study, quantitative data of the basins response to pressure (p) are obtained evaluating volume basins (v) at different pressures according to the previous computed EOS of the corresponding crystal. Using standard analytical EOS functions as the Vinet one,¹⁸ fittings to the sets of (v , p) points provide EOS parameters for each basin. In this way, we compute zero pressure basin volumes (v_i) and basin compressibilities (k_i). It now becomes interesting to find out that the macroscopic compressibility (k) of the crystal can also be expressed as a sum of contributions from these chemically meaningful basins:

$$\kappa = -\frac{1}{V} \frac{\partial V}{\partial p} = \sum_i f_i \kappa_i \quad (1)$$

where $f_i = v_i/V$ is the fractional occupation volume of the i -th basin. This equation holds if we keep the same formal definition for the basin compressibility as for the bulk, $\kappa_i = -(1/v_i)(\partial v_i/\partial p)$, and we apply the volume additive equation $V = \sum_i v_i$.^{19,20} Here V and v_i refers to a formula unit in the cell. Thus, not only basin volumes and charge populations are additive but compressibilities too, which leads to the potential rationalization of the macroscopic response of solids to pressure in quantum chemical contributions.

3. Differential Behavior of Electron Pairs under Pressure

Let us consider the simple schemes of bonding patterns present in solids as depicted in Figure 1. The universal binding energy relation (UBER) model^{21–23} establishes correlations between the electron density, bond strength, and bulk modulus of materials (see also ref 3). Considering these works, we can assume that regions with low average electron density are the natural candidates to be more affected by pressure. According to this chemical knowledge (and also to our quantitative analysis below), these regions can be clearly identified with arrows in each of the bonding schemes of Figure 1. If we move now to the ELF 1D profiles (Figure 2), it is found that, in all cases, these regions correspond to the basins with lower ELF values and, more specifically, the zones of these basins with the steepest ELF descent.

It is worth describing in detail these ELF profiles, and we take as a first example that of the diatomic N_2 molecular crystal (see Figure 2b). We only refer to the left-half of the plot since the curve is symmetric with respect to $R(N_2 - N_2) = 0.5$, where R is a reduced length coordinate. The two highest maxima are associated with the K-shells ($1s^2$ electrons) of the two N atoms. Between these two peaks, there is another peak identifying the highly localized electrons corresponding to the N_2 triple bond. For our purposes, the most interesting features appear in the regions of the ELF profile the arrows are pointed to. These are the basins of the lone pair electrons, which spread out toward the other N_2 molecule. A first-order saddle point is found between them at $R = 0.5$, and the ELF value at this point is negligible. This reflects the description of the solid with strongly orthogonalized molecular-like (N_2) wave functions,²⁴ which results in the absence of charge flow between N_2 molecules as pressure is applied: Pauli prevents the occurrence of intermolecular overlaps. The same happens in the mono-atomic Ne crystal and in between Na^+ and Cl^- in NaCl.

For all these situations, we have found an interesting result: only the outermost part of the pressure active basins, starting at the most external maximum of the ELF profile, is compressed. This can be interpreted as the result of the Pauli principle that prevents electron mixing with the core as well as with other units. Hence, on the one hand, a minimal distance between nonorthogonalized orbitals is preserved, and the compression does not affect the ELF well. On the other hand, the core structure is also maintained so that the only

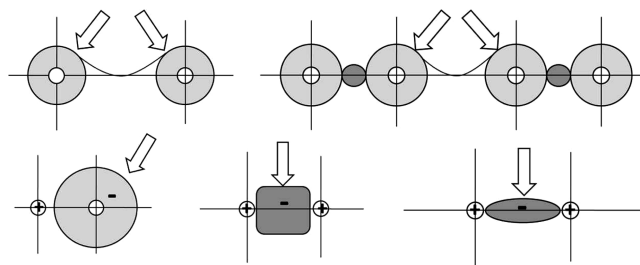


Figure 1. Bonding patterns in solids. From left to right and up to down: rare-gas monatomic, molecular diatomic, ionic, covalent, and metal. Arrows point to pressure sensitive basins. Dark regions stand for shared electrons, lighter ones for unshared electrons, and white ones for core electrons.

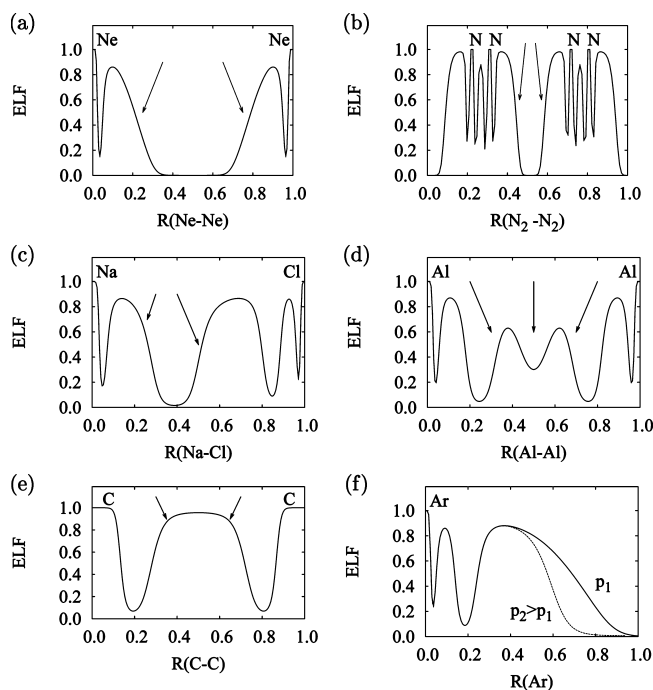


Figure 2. ELF profiles along the bonding path in solids: (a) rare-gas monatomic (Ne), (b) molecular diatomic (N_2), (c) ionic (NaCl), (d) metal (Al), and (e) covalent (C-diamond). ELF profiles, (f), for solid Ar at two pressures (p_1 and p_2). Arrows point at pressure sensitive basins. R is a reduced (relative) coordinate.

part of the atom that can subsume the pressure is the localized valence. Furthermore, due to the monosynaptic (they belong to only one nucleus) nature of these basins, the only part that is affected is the tail, as illustrated in Figure 2f, where the 1D ELF profile for Ar is plotted at two different pressures. The valence electron density in these closed shell systems surrounds the core with a radial distribution that is not at all uniform. The average electron density is much higher in the inner region of the lone pair basin than in its tail, and therefore, it is this part of the basin which is essentially affected by pressure. In addition, we should bear in mind that these closed-shell and lone pair basins contain electrons of the same spin, and again, Pauli repulsion makes their compression difficult.

As regards metals and covalent solids, the corresponding 1D ELF profiles clearly illustrate the electron shell structure of the atoms (high ELF values for the K-shell in C-diamond

Table 1. Zero Pressure Properties of Bulk and Basins of the Crystalline Phases for Most of the Elements with Atomic Numbers up to $Z = 20$ According to Our Calculations^a

E	SG	V_0	B_0	core		valence		bond		lone pair		$B_{0,i}$
				ρ_i	f_i	ρ_i	f_i	ρ_i	f_i	ρ_i	f_i	
He	$Im\bar{3}m$	34.71	0	—	—	0.0085	1.000	—	—	—	—	—
Li	$Im\bar{3}m$	20.37	13.6	0.302	0.049	—	—	0.007	0.951	—	—	13.5
Be	$P6_3/mmc$	7.91	121.0	0.477	0.081	—	—	0.039	0.919	—	—	108.4
B	$R\bar{3}m$	7.32	200.0 ^b	0.326	0.172	—	—	0.055	0.828	—	—	—
C	$Fd\bar{3}m$	5.71	442.8	2.637	0.021	—	—	0.104	0.979	—	—	445.4
N ₂	$Pa\bar{3}$	55.21	62.4	4.634	0.002	—	—	0.066	0.142	0.020	0.855	54.8
Ne	$Fm\bar{3}m$	24.26	3.1	18.819	0.001	0.048	0.999	—	—	—	—	4.9
Na	$Im\bar{3}m$	37.10	7.0	0.470	0.083	—	—	0.005	0.917	—	—	8.5
Mg	$P6_3/mmc$	22.84	36.1	0.684	0.094	—	—	0.015	0.905	—	—	32.6
Al	$Pm\bar{3}m$	20.41	57.5	0.955	0.073	—	—	0.026	0.927	—	—	52.7
Si	$Fd\bar{3}m$	19.72	97.1	1.417	0.053	—	—	0.031	0.947	—	—	92.3
P	$Cmca$	18.98	58.9	2.061	0.026	—	—	0.027	0.974	—	—	58.6
Cl ₂	$Cmca$	45.55	24.5	4.519	0.029	—	—	0.125	0.075	0.032	0.897	26.6
Ar	$Fm\bar{3}m$	39.37	0.7	6.033	0.006	0.030	0.994	—	—	—	—	0.3
K	$Im\bar{3}m$	70.88	4.7	0.150	0.238	—	—	0.002	0.762	—	—	4.5
Ca	$Fm\bar{3}m$	42.66	17.6	0.272	0.232	—	—	0.008	0.768	—	—	17.1

^a E and SG stand for elements and space group, respectively. Volume (V_0) is in Å³, bulk modulus for crystals (B_0) and for the basin with lowest ρ_i ($B_{0,i}$) are in GPa, average electron density (ρ_i) is in number of electrons per bohr,³ and volume occupation fraction (f_i).
^b Approximated from theoretical results of ref 25.

and for the K- and L-shells in Al) with bonding basins between them marked with the *pressure* arrows. In both cases, the valence basins are shared between two (or more) atoms so that both the ELF and density have high values along the bonding direction and both contours of the basin are subject to compression. Furthermore, since these basins feel several strong potentials from neighboring atoms, they are confined to smaller regions of space that, obviously, lead to smaller compressibilities. The high ELF values of the bonding basin in C-diamond reveal electrons highly localized with directional C–C bonds, whereas in Al, lower ELF values are associated with electrons delocalized among different bonding basins yielding a minimum in this 1D profile (see Figure 2d and e). Notice that the charge population, the average electron density, and the ELF values are lower in metals than in covalent bond basins since, in metallic elements, the number of valence electrons (up to three) is also lower, and the valence electrons are not completely paired. Moreover, the radial distribution of the electron density inside these basins is much more uniform than in the closed-shell systems, and we can think that all the electrons within the bonding basin are affected by pressure. The saddle point between the bond basin and the L-shell in Al has a low but not negligible ELF value that allows some charge flow between these basins as pressure is applied, leading to contributions from the outer core to the macroscopic compressibility. In fact, it is a common feature that the saddle point ELF value raises, and the ELF values in the bonding basin decrease with pressure, leading to a greater delocalization of the electron density.¹⁶ These features are in contrast with the static picture drawn from molecular and ionic crystals, where *bip* ELF values are low and the contribution of electron delocalization between basins is negligible.

We conclude that there exists two main classes of compounds as far as the effect of pressure is considered. On the one hand, metal and covalent crystals (from now on called *type A* solids) are mainly affected by strong forces.

On the other hand, molecular and rare-gas solids (from now on called *type B* solids) are characterized by the presence of electrons belonging to just one atom and are stabilized by weak interactions. Whereas in *type A* systems, the increase of interelectronic repulsions correlates with the material stiffness, *type B* solids can be seen as closed shell systems in which the underlying compression mechanism arises from the resistance to overlapping imposed by the Pauli principle.³ The fact that their cohesion energy, unlike in *type A* solids, increases with atomic size constitutes a natural proof of the different forces at work.

4. Chemical Contributions to Compressibility in Solid Elements of the s and p Blocks

Crystallographic as well as bulk and basin EOS parameters are collected in Table 1. This table also contains the average electron density and the volume occupation fraction of each of the basin types in which the corresponding unit cells of the solid elements of the s and p blocks are divided. Preliminary results for core, bond, and lone pair compressibilities of some elements have been published elsewhere.¹⁴ In Figure 3, we have used B_0 calculated values from Table 1 to plot the inverse of the bulk compressibility, along with the inverse of the compressibility of the basin more sensitive to pressure in each crystal.

The first question we would like to address is how the bulk compressibility compares with the compressibilities of the different basins that result from the ELF partition. It can be seen from eq 1 that two factors determine the contribution of the different basins to the total crystalline compressibility: on the one hand, the relative volume and, on the other hand, the basin compressibility. In all cases, and due to both factors, it is always possible to identify just one ELF valence basin for each crystal, as the basin that controls the pressure response of the crystal. This is apparent in Figure 3, where it is illustrated that not only the bulk modulus of each crystal,

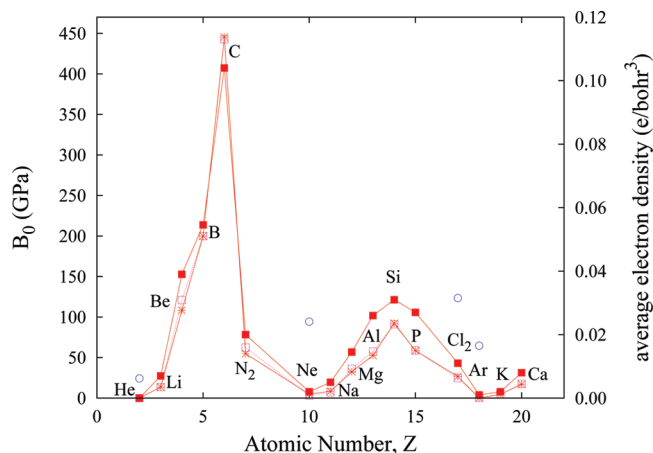


Figure 3. Dependence on the atomic number of bulk (empty squares) and valence basins (stars) compressibilities for most of the solid elements of the s and p blocks with $Z < 20$. Computed average electron densities of these basins (filled squares) are also included using the right y scale. Empty circles stand for uncorrected electron densities of He, Ne, Cl₂, and Ar.

and that of the *controlling* basin, follow the same periodic trend but also their absolute values are very similar. This reflects the fact that compressibility is a chemical property originated by valence electrons that are clearly identified by the topological partition of the ELF.

As we go from group I to VIII, we observe changes in the nature of that valence basin representing the main contribution to the bulk compressibility. It is first associated with a metallic bond basin (k_i in Li is $7.41 \times 10^{-2} \text{ GPa}^{-1}$, and the calculated crystal value is $7.35 \times 10^{-2} \text{ GPa}^{-1}$), then goes to a covalent bond basin (k_i in diamond C is $2.75 \times 10^{-3} \text{ GPa}^{-1}$ and k is $2.39 \times 10^{-3} \text{ GPa}^{-1}$), then to a lone pair basin (in N₂ k_i is $1.82 \times 10^{-2} \text{ GPa}^{-1}$ and k is $1.60 \times 10^{-2} \text{ GPa}^{-1}$), and ends with a closed-shell valence basin (in Ne k_i is $2.04 \times 10^{-1} \text{ GPa}^{-1}$, and the bulk value is $3.18 \times 10^{-1} \text{ GPa}^{-1}$). Under the ELF framework, this trend clearly recovers the decreasing connectivity of valence electrons along the periodic table. Metallic valences are in one extreme: in metals, the valence basins are located in the interstitial sites where they often form multicenter bonds; they are followed by prototypical two center two electron covalent bonds; lone pairs, which belong to just one center come next; and rare-gas solids, where all electrons are core-like, in the end. Overall, the contribution from the K-shell cores of these elements is almost negligible since their average electron density is very high, and strong forces are acting within them. Similar results are obtained from the analysis for the calculated elements of the second and third periods.

The correlation between the compressibilities of the crystals and the valence basins calls for a relationship similar to the ones derived in other theoretical and phenomenological models^{3,25} but involving, in our case, the average electron density of their respective valence basins. It is graphically illustrated in Figure 4, where a standard plot of the logarithms of the computed crystalline B_0 values versus the logarithms of the average electron densities of the *controlling* basins are displayed. Successfully, the computed average electron

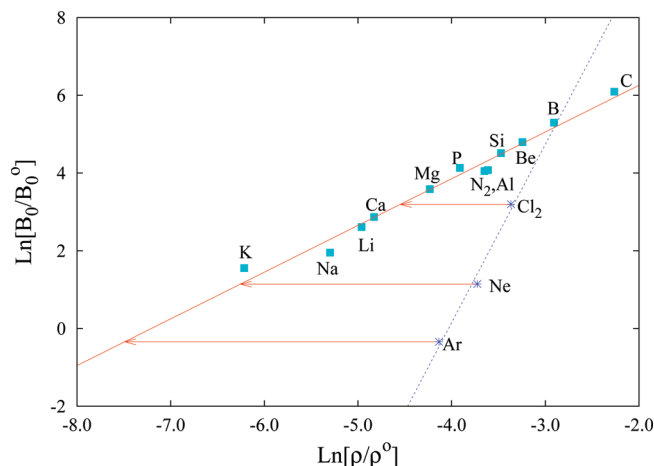


Figure 4. Linear fit showing the correlation between logarithms of zero pressure bulk modulus and logarithms of the average electron density of the *controlling* basin for *type A* solids. Missfitting to the general behavior is highlighted for *type B* solids with stars and a dotted line. B_0^0 (1 GPa) and ρ^0 (1 e/bohr³) allow to remove dimensions in the corresponding Ln functions.

density of the most sensitive ELF basin in each of all the studied *type A* crystals is very well described with a single linear function. The calculated slope (1.20) is very close to the value (1.26) reported by Gilman using the valence average electron density in elements of groups I–IV.³ Notice also that the reference state (B_0^0) in Gilman's plot is 1 Mbar. We should stress that, in our case, B_0 values come from the first-principles total energy calculations and the average electron densities from the partition induced by the ELF topology. It is remarkable that, in the pioneer work of Richards¹ with experimental data of more than forty elements, an equivalent plot to the one drawn here would have a slope of 1.25. A linear behavior (slope is equal to 1) is obtained for families of compounds under the so-called B_0 – V_0 inverse relationship proposed by D. L. Anderson and O. L. Anderson.²⁶ This kind of phenomenological expression has become very popular,²⁷ and identifies the formula unit volume as one decisive parameter to carry out correlations with the isotropic compression of solids. In our study, volume is subsumed in the average electron density of the *controlling* basin, which turns out to be a more general parameter to describe this behavior. Our results also reveal the ability of these valence basins more sensitive to pressure to enclose the periodic behavior of solid elements under pressure. Figure 3 shows the accuracy of this approximation, where the parameter ρ is able to recover the periodicity of the bulk modulus in *type A* solids.

As far as *type B* solids are concerned, when the electron density of their valence pressure active basins is displayed (empty circles in Figure 3), the periodic trend is broken. This behavior also appears in Figure 4, where the arrows indicate how far closed-shell crystals lay from the quasi-linear B_0 – ρ trend. Note that a different slope is expected for this *type B* solids (dotted line in Figure 4). As we have discussed in Section 3, the reason behind the different behavior exhibited by *type B* solids relies on the different character of the forces that are acting on their valence basins when pressure is

applied, which clearly do not follow a simple relationship with the electron density. *Type B* solids do not present basins with sharing electrons, and it is the average electron density of this kind of basins the one that correlates with B_0 . Nevertheless, it might be interesting to have for this *type B* solids an *effective* average electron density equivalent to the one observed in *type A* systems. If we introduce B_0 values of *type B* solids in the linear fitting expression obtained for *type A* solids, we obtain the *effective* average electron density for these systems, as pointed by the tips of the arrows in Figure 4. These *effective* electron densities for *type B* solids recover a correct picture regarding the periodical trend collected in Figure 3.

Graphical inspection of Figure 3 also reveals that the trends both along the periods and down the families are qualitatively reproduced. In fact, since the valence is conserved down the families, dropping the electronic saturation of the *controlling* basin becomes, in most cases, a reasonable approximation, so that a qualitative analysis could be made only in terms of the cell volume of the *controlling* basin (see Table 1). Given the fact that the volume increases as we go down, it is expected higher compressibilities for the heavier elements of the family. This is easily checked down the alkalines series Li–Na–K, where the volume increases from 20.4 to 70.9 Å³ and the bulk modulus decreases from 13.6 to 4.7 GPa. However, there are some cases where the cell volume turns out to be a deficient descriptor and resorting to the average electron density of the *controlling* basin becomes necessary in order to understand the exhibited periodicity. This is, for example, the case with Si and P. In spite of having nearly the same atomic volume, the greater number of valence electrons in Si results in a much lower compressibility. It should be noted that some general assumptions can still be made along the periods related to the nature of the compounds. The greatest volumes, along with the weak long-range forces, involved in both molecular crystals and rare-gas solids give rise to the most compressible solids. In the other extreme, compact covalent bond basins are the most difficult to compress, since they have the smallest *controlling* basins.

Finally, we would like to point out that the correlations we have shown between B_0 and the average electron density of valence basins have been computed after an exhaustive partition of the unit cell space into basins as induced by the ELF topology. Similar topological analysis under the framework of Bader's formalism yields valuable information in the case of closed shell systems, as alkali halides or oxide and selenide spinels,^{19,28} but it is not able to go inside the chemistry of the different types of bonding patterns shown in the systems explored here. In the case of many of the solid phases of these chemical elements, the average valence electron density can be computed from the difference between unit cell (and total number of electrons) and core volumes (and core electron populations). Core volumes in crystals are available for some elements from X-ray scattering experiments (see pages 116–117 in ref 3) and can be also estimated from ELF topological analysis in atoms under the reasonable assumption of negligible distortion of the inner electron shells when atoms go from a gas to a crystalline

phase. In fact, atomic core volumes and electron populations have been previously computed by Kohout and Savin²⁹ and Matito et al.³⁰ after much less expensive ELF topological analyses in atomic systems. We have checked that a good qualitative agreement with our results are obtained when the numbers from these calculations are used to display the periodic trends analyzed in the present study.

5. Conclusions

We would like to highlight the success of the underlying formulation behind our approach when looking for a comprehensive understanding of the crystalline response of elements to hydrostatic pressure. Our analysis provides a deeper insight into the periodicity of solids compressibility along periods and groups, stressing the relevance of the valence average electron density. ELF turns out to be an important tool due to its ability to identify the pressure *controlling basin* and to quantify its charge and volume. All the data calculated so far correlate with the chemical concepts as classically established: it is always the valence electrons involved in the bond, the lone pair or the closed-shell basins that are responsible for the main compression within the systems, whereas cores remain practically untouched. Our study does not only confirm these principles but also endows them with a theoretical background, enabling the ability to foresee future behavior. Metals and pure covalent solids belong to a different class of materials than that of the rare-gas solids and molecular crystals, as far as the origin of their response to hydrostatic pressure is concerned. The stiffness in the two former kinds of solids is dominated by repulsive electrostatic interactions among the valence electrons (from one to three in metals, four in covalent solids) that increase as pressure is applied. For rare-gas solids and molecular crystals, it has been shown that only the outermost part of the lone pair basin suffers compression under pressure. This region starts from the most external maximum of the ELF and reflects the lack of homogeneity of the electron density within the lone pair basins. Pauli repulsions within closed-shell basins and between closed-shells of different ions or atoms prevent the different constituents of these systems from coming closer under pressure.

Acknowledgment. Financial support from the Spanish MEC and FEDER programs under project MAT2006-13548-C02-02 and the Spanish MALTA-Consolider Ingenio-2010 project CSD2007-00045 are gratefully acknowledged.

References

- (1) Richards, T. W. *Proc. Natl. Acad. Sci. U.S.A.* **1915**, *1*, 411. Richards, T. W. *Proc. Natl. Acad. Sci. U.S.A.* **1923**, *9*, 73.
- (2) Bridgman, P. W. *Proc. Natl. Acad. Sci. U.S.A.* **1922**, *8*, 361.
- (3) Gilman, J. *Electronic Basis of the Strength of Materials*; Cambridge University Press: Cambridge, United Kingdom, 2003, pp 110–141.
- (4) Becke, A. D.; Edgecombe, K. E. *J. Chem. Phys.* **1990**, *92*, 5397.
- (5) Silvi, B.; Fourre, I.; Alikhani, E. *Monatshefte fur Chemie* **2005**, *136*, 855.

- (6) Kresse, G.; Furthmuller, J. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1996**, *54*, 11169.
- (7) Saunders, V. R.; Dovesi, R.; Roetti, C.; Causá, M.; Harrison, N. M.; Orlando, R.; Zicovich-Wilson, C. M. *CRYSTAL98 User's Manual*; University of Torino: Torino, Italy, 1998.
- (8) Noury, S.; Colonna, F.; Savin, A.; Silvi, B. *J. Mol. Struct.* **1998**, *450*, 59.
- (9) Perdew, J. P.; Zunger, A. *Phys. Rev.* **1981**, *23*, 5048.
- (10) Perdew, J. P.; Wang, Y. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1992**, *45*, 13244.
- (11) Towler, M., Crystal Resources Page, <http://www.tcm.phy.cam.ac.uk/~mtd26/crystal.html> (accessed March 10, 2008).
- (12) CRYSTAL Basis Sets Library, http://www.crystal.unito.it/Basis_Sets/ptable.html (accessed March 1, 2008).
- (13) Otero de la Roza, A.; Blanco, M. A.; Martín Pendás, A.; Luaña, V. *Comput. Phys. Commun.* **2009**, *180*, 157.
- (14) Contreras-García, J.; Martín Pendás, A.; Silvi, B.; Recio, J. M. *J. Phys. Chem. Solids* **2008**, *69*, 2204.
- (15) Contreras-García, J.; Martín Pendás, A. M.; Recio, J. M. *J. Phys. Chem. B* **2008**, *112*, 9787.
- (16) Contreras-García, J.; Silvi, B.; Martín Pendás, A. M.; Recio, J. M. *J. Phys. Chem. B* **2009**, *113*, 1068.
- (17) Contreras-García, J.; Silvi, B.; Martín Pendás, A.; Recio, J. M. *J. Chem. Theory Comput.* **2009**, *5*, 164.
- (18) Vinet, P.; Rose, J. H.; Ferrante, J.; Smith, J. R. *J. Phys.: Condens. Matter* **1989**, *1*, 1941.
- (19) Martín Pendás, A.; Costales, A.; Blanco, M. A.; Recio, J. M.; Luaña, V. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2000**, *62*, 13970.
- (20) Recio, J. M.; Franco, R.; Martín Pendás, A.; Blanco, M. A.; Pueyo, L.; Pandey, R. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2001**, *63*, 184101.
- (21) Banerjee, A.; Smith, J. R. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 6632.
- (22) Rose, J. H.; Ferrante, J.; Smith, J. R. *Phys. Rev. Lett.* **1981**, *47*, 675.
- (23) Martín Pendás, A.; Recio, J. M.; Francisco, E.; Luaña, V. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1997**, *56*, 3010.
- (24) Savin, A. *J. Chem. Sci.* **2005**, *117*, 473.
- (25) Wu, B. R.; Sung, C. M.; Lee, S.-L.; Tai, M. F. *Chin. J. Phys.* **2002**, *40*, 187.
- (26) Anderson, D. L.; Anderson, O. L. *J. Geophys. Res., [Atmos.]* **1970**, *75*, 3494.
- (27) Hazen, R. M.; Finger, L. W. *J. Geophys. Res., [Atmos.]* **1979**, *84*, 6723.
- (28) Waskowska, A.; Gerward, L.; Staun Olsen, J.; Feliz, M.; Llusar, R.; Gracia, L.; Marqués, M.; Recio, J. M. *J. Phys.: Condens. Matter* **2004**, *16*, 53.
- (29) Kohout, M.; Savin, A. *Int. J. Quantum Chem.* **1996**, *60*, 875.
- (30) Matito, E.; Silvi, B.; Duran, M.; Sola, M. *J. Chem. Phys.* **2006**, *125*, 024301.

CT900224W

Transferable Coarse Grain Nonbonded Interaction Model for Amino Acids

Russell DeVane,[†] Wataru Shinoda,[‡] Preston B. Moore,[§] and Michael L. Klein^{*†}

Center for Molecular Modeling and Department of Chemistry, University of Pennsylvania, 231 South 34th Street, Philadelphia, Pennsylvania 19104-6323, Research Institute for Computational Sciences (RICS), National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba Central 2, Umezono 1-1-1, Tsukuba, Ibaraki 305-8568, Japan, and Department of Chemistry and Biochemistry, University of the Sciences in Philadelphia, 600 S. 43rd Street, Philadelphia, Pennsylvania 19104

Received October 17, 2008

Abstract: The large quantity of protein sequences being generated from genomic data has greatly outpaced the throughput of experimental protein structure determining methods and consequently brought urgency to the need for accurate protein structure prediction tools. Reduced resolution, or coarse grained (CG) models, have become a mainstay in computational protein structure prediction performing among the best tools available. The quest for high quality generalized CG models presents an extremely challenging yet popular endeavor. To this point, a CG based interaction potential is presented here for the naturally occurring amino acids. In the present approach, three to four heavy atoms and associated hydrogens are condensed into a single CG site. The parametrization of the site–site interaction potential relies on experimental data thus providing a novel approach that is neither based on all-atom (AA) simulations nor experimental protein structural data. Specifically, intermolecular potentials, which are based on Lennard-Jones (LJ) style functional forms, are parametrized using thermodynamic data including surface tension and density. Using this approach, an amino acid potential data set has been developed for use in modeling peptides and proteins. The potential is evaluated here by comparing the solvent accessible surface area (SASA) to AA representations and ranking of protein decoy data sets provided by Decoys 'R' Us. The model is shown to perform very well compared to other existing prediction models for these properties.

I. Introduction

With the rate at which genomic data have become available over the last several years, the need to accurately predict protein structures from the amino acid sequence has become paramount. Protein structure prediction has employed many approaches to attempt this feat with a few noted here.^{1–16} Many of these models are based on a reduced resolution or

coarse grained (CG) representation of the protein.^{1–3,5,17–30}

Such models are used because they provide a means to expand the capabilities of existing computational resources. However, this enhanced efficiency comes necessarily at the cost of reduced resolution in the description of the system since typically the details of several atoms are essentially averaged into a single CG interaction site. When this reduction of resolution is introduced, it is crucial to accurately capture the phenomenon of interest for the specific system under investigation or the resulting CG description is simply a toy model capable of providing only generic behavior.

The history of CG models for amino acids dates back over three decades at least, with the pioneering work of Levitt

* Corresponding author e-mail: klein@lrsm.upenn.edu.

[†] University of Pennsylvania.

[‡] National Institute of Advanced Industrial Science and Technology (AIST).

[§] University of the Sciences in Philadelphia.

and Warshel, where they used reduced resolution models to investigate protein dynamics.¹ Since then, much work has been put forth to develop reduced resolution models to overcome the computational barriers of studying proteins. Much of this work has been built on knowledge based models which take advantage of the existing structural data acquired from experimental techniques including X-ray crystallography and NMR spectroscopy. Tanaka and Scheraga were the first to develop knowledge based contact potentials from the frequency of contacts between residues in known structures.² This approach was then extended by many groups to include for example distance and direction dependence.^{3,5,17–20} Others such as Huang et al. have broken from this paradigm to use an approach that does not rely on known protein structural data other than the general nature of the packing of hydrophobic and hydrophilic residues in protein structures.^{9,31}

Further, the use of CG protein models for molecular dynamics (MD) simulations has helped extend the capabilities of current computational resources allowing access to much larger temporal and spatial scales.^{32–47} In particular, the MARTINI model has become a widely used CG model with an array of applications to protein and membrane simulations.^{33–35} The later approach employs a LJ potential for nonbonded interactions with parameters in a tabulated form based on the interaction types (hydrophobic/hydrophobic, hydrophilic/hydrophobic, etc.). Many other groups have pushed CG models to a still higher level with approaches such as force matching^{38–40} and other schemes which have enjoyed success.^{41–53}

Herein we present the application of a recently proposed methodology to the development of CG nonbonded interaction potentials for amino acids.^{44,48–50} The current model is essentially a distance dependent potential which distinguishes the 20 naturally occurring residues to a high level such that only a few residues are modeled with the same parameters. Molecular simulations are used to parametrize CG sites that are used here in a nondynamical representation of proteins. Experimental thermodynamic data are used as the target for parametrization of the CG site interactions. This approach builds on previous work by Nielsen et al.⁵² who used a similar approach to obtain CG parameters for a series of alkanes and Shinoda et al.^{49,50} who extended the CG parameters to polyethylene glycol surfactants. Related approaches have been reported recently but with notable differences.^{54,55} The work by Basdevant et al.⁵⁴ uses a hybrid $1/r^6$ repulsive term and Gaussian attractive term that is parametrized using AA force fields. The work by Han et al.⁵⁵ uses a thermodynamic based parametrization approach but with a higher resolution mapping.

Herein, the current model is used to predict the native structures from several protein decoy sets with a level of success on par with knowledge based potentials. Finally, although not explicitly demonstrated here, the nonbonded CG model presented could be coupled with an appropriate intramolecular force field and used for MD simulations of peptides and proteins in a spirit similar to the MARTINI force field. Indeed, an initial application of the present approach was previously utilized in an MD simulation

investigation of peptide nanoring self-assembly.³² In Section II, the methods are discussed including the force field details and parametrization. Section III presents results and discussion including evaluation of the model through solvent accessible surface area (SASA) calculations and the ranking of protein decoys. Finally, Section IV closes with the conclusions.

II. Methods

A. Coarse Grain Model. Mainly for ease of implementation, the present CG model employs LJ style nonbonded potential functions.⁴⁹ For the models developed herein, all CG beads interact via a LJ(9-6) potential given in eq 1. The CG sites are charged, and this is also included in the electrostatic contribution given by eq 2

$$v(r)_{9-6} = \frac{27}{4}\epsilon\left(\frac{\sigma^9}{r^9} - \frac{\sigma^6}{r^6}\right) \quad (1)$$

$$v(r)_{elec} = \frac{1}{4\pi\epsilon_0} \frac{q_1q_2}{r} \quad (2)$$

The choice of prefactor for the LJ function is selected such that $v(\sigma) = 0$ and ϵ is the minimum energy. The choice of the LJ functional form (for example the 9–6 in eq 1) is essentially an adjustable parameter used in the fitting procedure. We have previously explored various options including 6–4, 8–4, 10–4, 9–6, and 12–4.⁴⁹ For alkane interactions the choice of the 9–6 functional form was validated by comparison of the CG liquid structure to AA MD results. The adoption of the 9–6 functional form reflects a desire to maintain consistency with existing CG models.^{49,50} A simple truncation, implemented at a distance of $rcut = 15 \text{ \AA}$, is employed for the long-range cutoff, with no smoothing or shifting. This length is sufficient to avoid gross artifacts resulting from the discontinuity; however, the cutoff length, $rcut$, does affect the thermodynamic properties and so is treated as a CG fitting parameter. The parameters in the LJ function are fixed by reproducing thermodynamic data. For bulk solutions, both ϵ and σ can be unambiguously fixed with a combination of density and surface tension. The cross interactions arising from nonidentical CG sites can be generated using the combining rules given by eqs 3 and 4

$$\sigma_{ab} = \frac{\sigma_{aa} + \sigma_{bb}}{2} \quad (3)$$

$$\epsilon_{ab} = \sqrt{\epsilon_{aa}\epsilon_{bb}} \quad (4)$$

Here, ϵ_{aa} and σ_{aa} represent the self-interaction values for ϵ and σ values, and ϵ_{AB} and σ_{AB} represent those for the cross interactions. Finally, the electrostatic interactions are calculated using an effective dielectric constant of 80 and without employing a cutoff for electrostatic interactions. The dielectric constant is given below in the form of scaled charges with the value of 0.1118.

Although the intramolecular force field is not discussed or used in detail here, it was necessary to employ bonds and angles in the development of the nonbonded interactions for PHE, TRP, and TYR (throughout the paper, the 3-letter

amino acid labels will be used) side chains. For those cases, the intramolecular interactions are modeled via a harmonic potentials given by eqs 5 and 6

$$V(r)_{bond} = K_b(r - r_0)^2 \quad (5)$$

$$V(r)_{angle} = K_a(\theta - \theta_0)^2 \quad (6)$$

where r_0 and k_b represent the equilibrium bond length and force constant for bonds, and θ_0 and k_a represent the equilibrium angle value and force constant for bends.

B. Molecular Dynamics Simulations. CG MD simulations were performed using the LAMMPS code developed at Sandia National Laboratory and extended by our group (now a part of the standard LAMMPS release) to implement our CG models.⁵⁶ An integration time step of up to at least 25 fs can be used to evaluate the nonbonded interactions. Intramolecular degrees of freedom have to be dealt with on a case by case basis as the force constant will determine the time step necessary. For these cases, the multitime step integrator rRESPA can be used to separate the various degrees of freedom.⁵⁷

Electrostatic interactions were calculated using the Particle–Particle Particle–Mesh method implemented in the LAMMPS MD code.^{58–61} Nonbonded interactions are excluded for bonded CG sites separated by two or less bonds (1–2, 1–3). Interactions between CG sites separated by 3 or more bonds are included without scaling. For the sake of evaluating these exclusions, the connectivity (bonds) of the model was considered although the bond energy was not calculated in the potential evaluation. Only the nonbonded van der Waals and electrostatic energy were used. The bulk system MD simulations included roughly 500 CG sites each. For surface tension simulations, roughly 50 ns simulation time was used. Systems were set up by equilibrating with an isothermal–isobaric (NPT) simulation. The equilibrated system box dimension was then extended in the z-direction creating a vacuum region large enough such that no interactions extended through the vacuum via the periodic boundary conditions to the opposite side. The extended system was then used in a MD run in a canonical (NVT) simulation. The surface tension, γ , was then calculated via

$$\gamma = \frac{L_z}{2} \left[P_{zz} - \frac{(P_{xx} + P_{yy})}{2} \right] \quad (7)$$

where L_z is the box dimension in the z-direction, and P_{aa} is the isotropic pressure tensor for each Cartesian direction where $a = x, y, z$.⁶² Density calculations were performed with NPT simulations averaged over roughly 2 ns simulation time. Finally, all MD simulations were performed at a temperature of 303.15 K, unless otherwise specified.

C. Solvent Accessible Surface Area. One property of interest in proteins is the SASA which is proportional to protein solvation free energy and represents an important property in protein research and design. In the work of Lee and Richards, it was shown that the hydrophobic residues tend to have a larger reduction of SASA when going from extended chains to the native conformation compared to the polar residues.⁶³ Succinctly, the SASA is the area of the

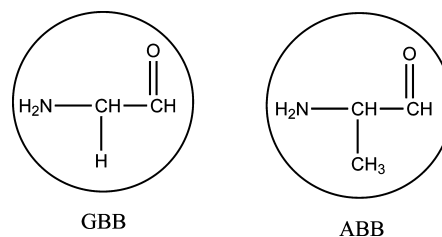


Figure 1. Shown here are the CG mapping for the backbone sites. The standard backbone CG site (GBB) is shown on the left, and the ALA CG site (ABB) is shown on the right.

surface traced out by rolling a probe sphere, representing the solvent, over the surface of the protein. Any area accessible to the probe that does not require the probe to overlap with neighboring atoms is considered solvent accessible. The SASA was calculated here via the implementation in VMD.⁶⁴ For the AA level calculations, a probe radius of 1.4 Å and VMD default atomic radii for the protein atoms were used. For the CG SASA calculations, a probe radius of 2.5 Å was used. This value was calculated from the minimum of the potential energy function, r , where $V(r)/r = 0$, for the CG water–water interaction. CG site radii for the amino acids were calculated in a similar fashion. Note that the minima of the LJ functional forms used herein are not equal to the standard LJ (12–6) minimum of $r = 2^{1/6}\sigma$ but are given by $r = 1.5^{1/3}\sigma$ and $r = 3^{1/8}\sigma$ for the LJ (9–6) and LJ (12–4), respectively.

D. Parameterization. In the CG model presented here, the mapping includes roughly three to four heavy atoms and adjacent hydrogens per CG site. Each amino acid is divided into a backbone and side chain section. The mappings for all of the naturally occurring amino acids are given in Figures 1, 2, and 3. All amino acids in this model (including GLY) share the same backbone representation with the exception of ALA. The standard backbone CG site includes the carbonyl carbon and oxygen, nitrogen, α carbon, and related hydrogens. GLY is of course represented by the single standard backbone CG site. Due to its small side chain, the ALA amino acid is mapped to a single site containing the backbone and side chain. The mappings for the backbone sites are shown in Figure 1 with the standard CG backbone site (labeled GBB in Table 3 used for all amino acids except ALA on the left) and the ALA CG model (labeled ABB in Table 3 given on the right).

Most side chains are represented by a single CG site. The exceptions to this are the PHE, TYR, LYS, and ARG side chains, which are modeled with two CG sites, and TRP which is modeled with three. Finally, Figures 2 and 3 show the mapping for all of the neutral and charged side chains, respectively.

The side chain analogue molecules used in the model development of the neutral amino acids are given in Table 1. To parametrize the LJ9-6 potentials for the self-interactions of the CG sites (SER-SER, LEU-LEU, etc.) the surface tension and density were reproduced in bulk solutions of the representative side chain analogue where possible. For example, this was not possible for the TRP side chain analogue 3-methylindole due to the complexity (too many parameters to fix simultaneously with only two observables)

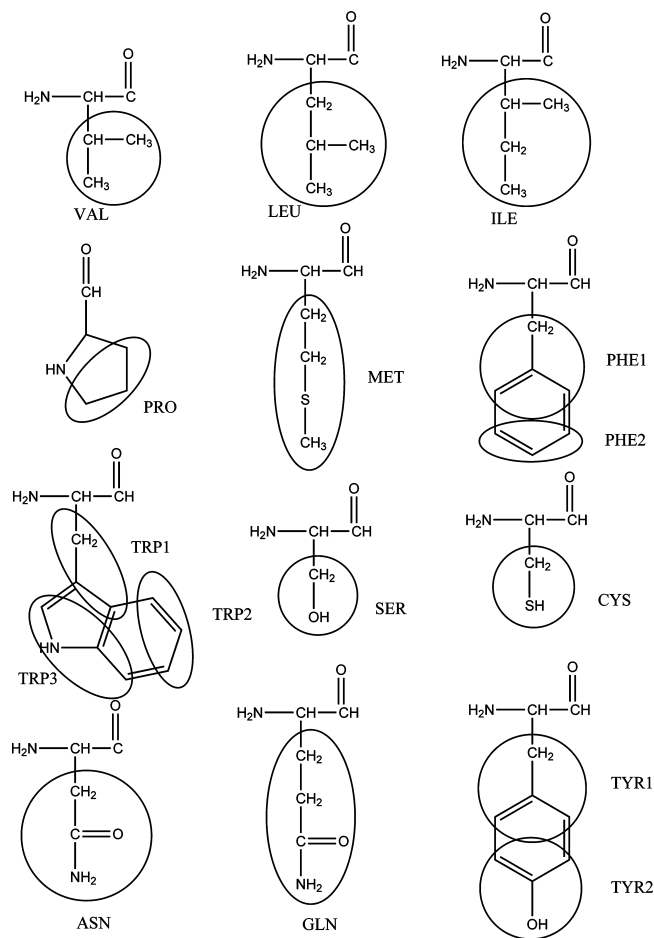


Figure 2. Shown are the mappings for the amino acid side chains. The label here corresponds to the labels used in the interaction potential Table 3.

and the fact that it is a solid at room temperature with a melting point of around 370 K. Instead, the CG sites for this side chain were developed in a stepwise process (see below). On the other hand, having slightly higher melting point did not alter the approach for ASN and GLN side chains. A slightly higher temperature was used in the simulations for the parametrization of this CG sites. Strictly speaking, the CG models are parametrized at a specific temperature and thus not transferable to different temperatures; however, there is a small range of temperature which will not alter the properties to a large degree. Results for the surface tension and density of each side chain analogue is compared to experimental values in Table 2.⁶⁵ The parameters for the VAL side chain are from previous work.⁴⁹ The parameters for LEU and ILE were developed using the analogues 2,5-dimethylhexane and 3,4-dimethylhexane, respectively, since the true analogues are gases at room temperature. Equilibrium bond lengths of 3.87 Å and 3.22 Å with force constant of 5.0 kcal/mol were used for each, respectively.

For PHE, TYR, and TRP, it was necessary to implement a bond between the CG sites in order to parametrize the overall CG assembly. For PHE and TRP the chosen minimum energy bond lengths, 2.5 Å, and the force constant of 150 kcal/mol were used for all bonds. For TYR, the minimum energy bond length, 2.9 Å, and the force constant

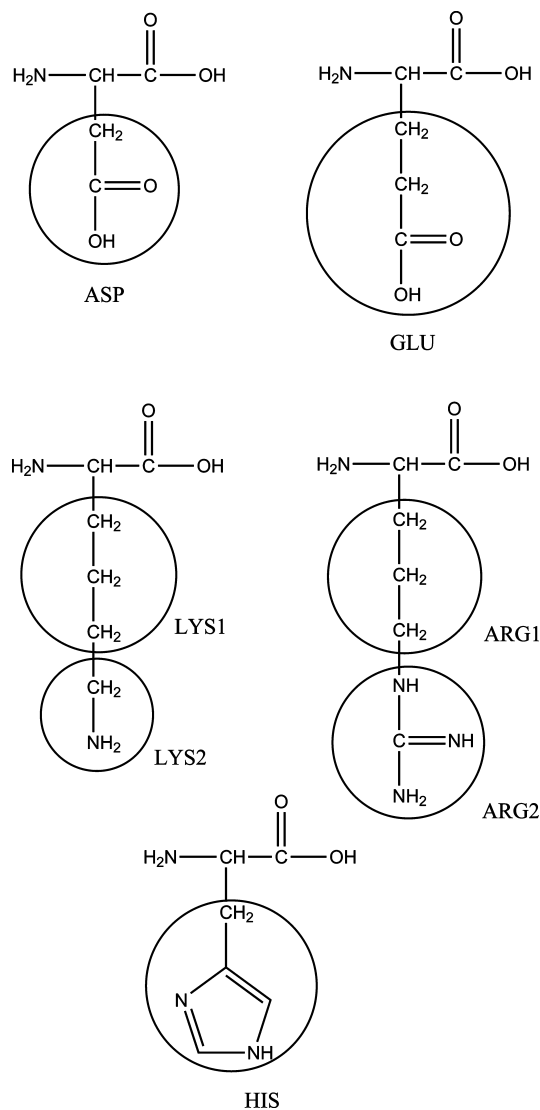


Figure 3. The CG mappings for the charged residues are shown here. ASP, GLU, LYS2, and ARG2 are all modeled with the same interaction parameters. LYS1 and LYS2 are modeled with the VAL CG site parameters. Note that HIS is included in this figure although it is not modeled with a charge in the current model.

Table 1. Side Chain Analogue Molecules

residue	analog	residue	analog	residue	analog
ALA	none	MET	methyl-ethyl sulfide	THR	ethanol
VAL	propane	PHE	toluene	CYS	methanethiol
LEU	isobutane	TRP	3-methylindole	ASN	acetamide
ILE	butane	GLY	none	GLN	propionamide
PRO	propane	SER	methanol	TYR	cresol

of 150 kcal/mol were used. The choice of a two-site CG model for PHE and TYR was used to maintain consistency with the mapping of roughly 3 to 4 heavy atoms per CG site. This is admittedly a gross overestimation of these side chains and the implications arising are discussed below. The PHE side chain was developed from the CG sites for p-xylene (XYL) and benzene (BEN). Symmetric two site models were made for each of these two molecules and parametrized to reproduce the respective surface tension and density. These sites were then combined to produce the PHE

Table 2. Thermodynamic Data from CG Simulations Compared with Experiment

molecule	surface tension (mN/m)		density (g/cm ³)	
	simulation	experiment	simulation	experiment
VAL (PRO) (hexane)	17.0±0.5	17.4	0.65	0.65
LEU (2,5-dimethylhexane)	19.0±0.5	18.8	0.69	0.69
ILE (3,4-dimethylhexane)	20.3±0.5	20.7	0.72	0.71
MET	24.0±0.5	23.2	0.83	0.83
PHE	26.7±0.5	27.3	0.86	0.86
SER	22.0±0.5	22.9	0.78	0.78
THR	22.7±0.5	22.7	0.78	0.78
CYS	22.0±0.5	23.0	0.86	0.86
ASN ^a	37.5±0.5	39.4	1.04	1.00
GLN ^a	30.7±0.5	30.6	0.95	0.93
TYR ^b	35.1±0.5	35.1	1.00	1.03

^a 358.0 K. ^b 315.5 K.

Table 3

side chain	LJ9-6 parameters		radii	
	ϵ (kcal/mol)	σ (Å)	CG (Å)	PDB (Å)
VAL	0.469	4.59	2.63	2.35
LEU	0.536	5.06	2.90	2.6
ILE	0.487	5.18	2.96	2.6
PRO	0.469	4.59	2.63	2.34
MET	0.835	5.03	2.88	2.71
PH1	0.435	4.91	2.81	2.57
PH2	0.413	4.33	2.48	2.3
TR1	0.473	4.41	2.52	2.33
TR2	0.413	4.33	2.48	2.3
TR3	0.753	4.66	2.67	2.51
SER	0.580	3.68	2.11	1.96
THR	0.666	4.24	2.43	2.27
CYS	0.656	4.16 (2.40)	2.38	2.18
ASN	0.870	4.15	2.38	2.42
GLN	1.192	4.74	2.71	2.66
TYR1	0.435	4.91	2.81	2.57
TYR2	0.700	4.10	2.35	2.48
ASP	0.497	4.00	2.29	2.4
GLU	0.497	4.00	2.29	2.64
LYS1	0.469	4.59	2.63	2.33
LYS2	0.497	4.00	2.29	1.91
ARG1	0.469	4.59	2.63	2.34
ARG2	0.497	4.00	2.29	2.32
HIS	1.400	5.40	3.09	2.73
GBB	0.870	4.15	2.38	2.4
ABB	1.200	4.74	2.71	2.69

side chain model. Figure 4 shows the CG sites and corresponding AA representations (without hydrogens for clarity) used in the stepwise process. The blue and red spheres represent the BEN and XYL CG sites, respectively. In the PHE side chain CG model (far right of Figure 4), the XYL CG site, represented by the red sphere, is labeled PHE1, and the BEN CG site, represented by the blue sphere, is labeled PHE2 in the interaction database (see Figure 4). The combination rules, eqs 3 and 4, from the Methods section were used to develop the XYL-BEN cross interaction. The result of using this approach can be seen in the reproduction of the bulk properties for the PHE side chain model in Table 2. The XYL CG site was then combined with a new site to create the cresol (TYR side chain) CG model in a similar fashion. The labels in the TYR CG model are TYR1, modeled with the XYL CG site and TYR2 modeled with the new -OH containing cresol CG site (see Figure 4).

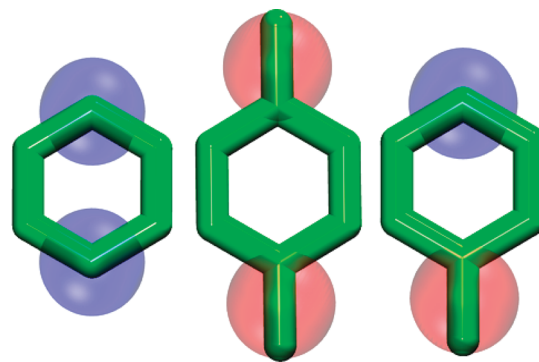


Figure 4. The models used to develop the phenyl based rings (PHE, TYR, and TRP) are shown here. The CG model for benzene (left and blue) and p-xylene (middle and red) were combined to produce the PHE CG model (right). These sites were also used in the TYR and TRP CG models as described in the text.

Finally, the benzene CG site was used to develop a CG model for the TRP side chain.

The analogue molecule 3-methylindole was used to represent the TRP side chain. The CG mapping for the TRP side chain is shown in Figure 2. The new CG sites TRP1 and TRP3 represent propene and allylamine, respectively. Each of these molecules is represented by a single distinct CG site that was parametrized in a similar fashion to the amino acids with side chains represented by single CG sites. In a stepwise process the TRP side chain was built up first by combining the BEN and ALL CG sites to create the CG model for aniline. The bond was modeled with the equilibrium bond length and force constant stated above. This process yielded the BEN-ALL cross interaction from the bulk aniline simulations. Then, the BEN and ALL beads were combined with the PPE CG site to create the TRP side chain CG model. The combination rules were employed to fix the interactions BEN-PPE and ALL-PPE.

For the charged residues (ASP, GLU, LYS, and ARG) a primitive model was developed including charges. Figure 3 shows the CG mapping for the charged residues. It should be noted that although HIS is included here, it is not modeled with a charge in the current model. The CG sites labeled ASP, GLU, LYS2, and ARG2 are identical in the current model. These sites were parametrized to reproduce the surface tension of guanidium salt solutions at various concentrations. A charge of 0.1118 (with $\epsilon_0 = 1$ in eq 2 scaled down from unity to compensate for the solvent screening that would take place normally but is absent due to the CG water model being represented by a single chargeless nonpolar spherical site) is included on the sites ASP, GLU, LYS2, and ARG2. All other sites are modeled without charge. The addition of a scaled charge is part of a recent development designed for modeling charged surfactants. Details will be part of a forthcoming manuscript. As stated above, the current model will not distinguish ASP from GLU or LYS from ARG. For LYS and ARG, the charged CG sites, labeled LYS2 and ARG2 (modeled with a charge), are combined with a CG site equivalent to the VAL side chain CG site (and modeled neutral) and labeled LYS1 and ARG1,

Table 4. Comparison of SASA for CG and AA Models^a

PDB I.D.	CG (Å ²)	AA (Å ²)	ratio (CG/AA)
lysozyme (2vb1)	6479.4	6676.4	0.97
myoglobin (2jho)	8040.1	8222.4	0.98
ribonuclease (1cv9)	7085.9	6968.8	1.02
flavodoxin (1flv)	7539.8	7713.3	0.98

^a The SASA was calculated for the CG and AA representations with a probe radius of 2.5 Å and 1.4 Å, respectively.

respectively in Figure 3, to create the two bead CG site for these side chains.

III. Results and Discussions

A. Surface Area. Evaluation of force fields for complex molecules such as proteins is a difficult task. Indeed it is an ongoing process where many systems must be explored in order to provide a good representation of possible interactions that can arise. Nonetheless, here we set out to provide at least a preliminary level of evaluation. For protein–protein interactions, steric effects arising from side chain packing can play an important role in determining the spatial arrangements. Therefore the relative sizes of amino acid side chains is a key property to preserve in any model of amino acids. Further, it is understood that CG modeling often distorts the representation of molecules especially when mapping to spherical sites. To evaluate the effects of this mapping, the SASA was calculated for each amino acid side chain and backbone unit from a set of PDB databank structures with less than 30% homology. For these calculations, each side chain or backbone unit was calculated as a solitary group of atoms (excluding all other atoms in the structure) to give a SASA for that residue side chain or backbone type (GLY or ALA) taking into account only the various conformations of each subunit. From these calculations, the radii were calculated for each subunit and are shown in Table 3. The results are compared to the radii calculated for the CG model side chain units as described in the Methods section. As shown in Table 3, the values compare well suggesting conservation of SASA with the mapping from AA to CG. It is important to note that the radii for the CG model were not based on all atom or PDB calculations but arise from the parametrization process as described above. To further evaluate the SASA predictive ability of the CG model, the SASA for full proteins were calculated. Table 4 shows the results of the SASA calculations for several proteins comparing the CG results to those from the AA level. The results show that despite the fact that the CG model uses a bulky solvent (combining 3 water molecules) the SASA is consistent in the CG and AA models. Finally, it should be mentioned that calculation of SASA is not straightforward, and many things, such as the probe radius, can affect the results. Measurements made here were done as consistently as possible using standard techniques and radii for the probe and atomic radii. Therefore, it is assumed that this is at least a reasonable comparison of the SASA.

B. Predicting Native Structures. The ranking of structures in decoy sets such as those provided by Decoys R Us is a useful test for evaluating amino acid potentials and has

become a somewhat standard evaluation tool.⁶⁶ The decoy sets are composed of native protein structures with decoys for each native structure generated through various techniques. The primary use of the protein decoy sets is to test a model's ability to distinguish the native structure from the non-native decoy structures. Here we evaluated the current CG model using five decoy sets (*4state_reduced*, *fisa*, *fisa_casp3*, *lmds*, *lattice_ssfit*) provided by Decoys R Us (<http://dd.stanford.edu>).^{14,66–69} For comparison, the MARTINI force field was also used to rank the decoy sets.³⁵ For each structure, the CG model was mapped onto the AA model by placing the CG sites at the center of mass of the representative heavy atoms. The nonbonded potential presented here (and the MARTINI nonbonded potential) was then used to calculate the potential energy of each structure. These energies were then ranked from lowest to highest. The observable is the ranking of the potential energy of the native structure compared to the decoys. A value of 1 indicates that the native structure was predicted to have the lowest energy (is the most stable) of the set. The higher the value the worse the ability of the model to identify the native structure from the decoys. The ranking of the native structures for each set is shown in Table 5. The results for the model presented here are shown in the columns labeled CG, and those for the MARTINI model are labeled MARTINI.³⁵

During the evaluation of the decoy sets, one issue had to be addressed in an ad-hoc fashion. Disulfide bonds arising between CYS residues had to be treated as a special case. The parametrization the CYS side chain site used bulk methylmercaptan solutions leading to an effective size which is much larger than the center of mass distance that arises when the CYS residues are involved in a covalent bond. To address this issue, sigma for the CYS-CYS interaction was shortened to a value of 2.4 Å to represent the CYS disulfide bond. However, it is important to note that the cross terms (CYS-LEU, CYS-SER, etc.) generated with the combination rules used the original bulk solution sigma value of 4.16 Å. To test the effect of the CYS-CYS interactions, the structures were re-evaluated with all of the CYS-CYS interactions removed. The results did not show significant differences which indicates that the use of the modified sigma value is reasonable for the structures evaluated here. However, to make a more general potential, this issue will have to be dealt with and the CYS-CYS will need to be categorized as either covalent or noncovalent. Typically CYS-CYS interactions involved in a disulfide bond would be excluded from the nonbonded interactions; however, the current setup does not take into account these bonds and thus they had to be dealt with with a reduced sigma value. For the MARTINI model, the calculation was repeated with all of the CYS-CYS interactions turned off and showed no significant difference in the results.

The results in Table 5 show that the potential is extremely effective at picking the native structure from the decoy sets. The present model ranked the native structure as the most stable in 68% of the decoy sets and ranked the native structure as one of the top 3 most stable structures in 77% of the decoy sets. For the sake of comparison, other methods are shown including MARTINI (mentioned above and not

Table 5. Results of Ranking Native Structures for Decoy Sets from Decoys 'R' Us^a

	DFIRE	Rosetta	ModPipe-Pair	ModPipe-Comb	DOPE	MARTINI	CG	Z-score	% NEG
<i>4state_reduced</i>									
1ctf	1/3.86	1	1	1	1	2	1	1.99	99.2
1sn3	1/3.79	1	1	1	1	398	29	1.39	99.4
2cro	1/3.29	5	1	1	1	397	1	2.47	99.0
3icb	4/2.28	6	151	8	1	308	282	0.31	99.4
4pti	1/3.62	1	1	1	1	254	267	0.35	99.6
4rxn	1/3.33	1	6	1	1	613	227	0.46	99.6
<i>fisa</i>									
1fc2	254/0.23	158	491	453	375	458	32	1.28	99.6
1hdd-C	1/4.50	90	293	135	1	131	3	1.95	98.2
2cro	1/6.33	26	11	19	1	58	1	4.22	98.8
4icb	1/6.91	1	196	167	1	482	1	4.20	95.4
<i>fisa_casp3</i>									
1bg8-A	1/5.35	1068	1	282	1	1200	1	4.35	4.6
1bl0	1/4.50	960	4	86	1	972	1	NA	0.1
1eh2	NA	NA	NA	NA	NA	2413	2	0.31	0.2
1jwe	1/6.26	1177	1	6	1	1407	1	2.37	0.6
smd3	NA	NA	NA	NA	NA	1	1	1.86	0.5
<i>lmds</i>									
1ctf	1/3.54	1	1	1	1	108	1	2.01	99.8
1dtk	1/2.62	9	4	1	1	197	154	-0.57	98.2
1fc2	501/-5.72	291	325	222	476	2	1	4.55	92.6
1igd	1/5.16	1	1	1	1	1	1	2.54	96.8
1shf-A	1/6.68	5	24	7	1	1	1	3.16	99.5
2cro	1/4.70	2	4	12	1	1	1	6.14	100
2ovo	1/3.21	29	5	2	1	346	34	1.18	100
<i>lattice_ssfit</i>									
1beo	1/12.09	1	1	1	1	1602	1	2.97	53.9
1ctf	1/10.05	1	1	1	1	1	1	3.29	70.9
1dkt-A	1/6.87	1	1	1	1	6	3	2.88	10.2
1fca	1/7.18	1	1	1	1	8	1	2.42	90.8
1nkl	1/9.29	1	1	1	1	1086	1	3.04	26.2
1pgb	1/11.87	1	1	1	1	186	1	2.71	68.2
1trl-A	1/6.32	45	1	1	1	7	1	2.78	69.4
4icb	1/7.81	1	1	1	1	2	1	3.07	35.5
total correct	26	14	17	17	27	5	21	NA	NA

^a The lower the value, the better the performance of the model. A value of "1" indicates that the model predicts the native structure to be the most stable of all of the decoy structures. The column labeled "MARTINI" shows the results from the MARTINI model.³⁵ The column heading "CG" represents the results from the current CG model. All other results were taken from Zhou and Zhou^{73,74} and Shen and Sali.¹³ The Z-score values and the percentage of structures reporting a negative potential energy value with the current model are shown under the "Z-score" and "%NEG" heading, respectively. For comparison, the ranking and Z-scores have been included for the "DFIRE" results (ranking/Z-score).

designed for this type of evaluation), DFIRE1, Rosetta, ModPipe-Pair, ModPipe-Comb, and DOPE.^{14,67,70-75} The values for these methods were taken directly from the references of Zhou and Zhou^{73,74} and Shen and Sali.¹³ Another evaluation of the quality of ranking is the Z-score which indicates the models discriminatory ability. We define the Z-score as

$$Z = \frac{\langle E \rangle - E_N}{\sigma} \quad (8)$$

where E_N is the energy of the native structure, $\langle E \rangle$ is the average energy of all structures in a given set (for example *4state_reduced: 1ctf*), and σ is the standard deviation of the distribution of energy values for a given set. The Z-scores for each set are given in Table 5. To prevent the Z-scores from being skewed by structures with unreasonably high energies, any decoy structure with a positive potential energy value was not included in the Z-score calculation. The percentage of structures that had negative potential energy values as determined by the current model are listed in Table

5 under the heading "%NEG". It should be noted that the *fisa_casp3* set had a high percentage of decoy structures with a positive potential energy according to the current model, and so the Z-scores from those sets must be judged accordingly. Further, the *lattice_ssfit* had a modest amount of structures evaluated with positive potential energies. Finally, all of the native structures had a negative potential energy. For the sets in which the native structure was ranked as the most stable structure, the average Z-score was 3.13. For comparison, Z-scores for the DFIRE model have been included in Table 5. The average Z-score from the DFIRE model for all of the systems tested here was 5.21.

The *4state_reduced* structures 3icb, 4pti, 4rxn, and the 1dtk structure from the *lmds* set are ranked very poorly by the current model. However, closer examination of these structures indicates that a substantial portion of the bad contacts arise from interactions involving phenyl based residues. The problem results from the mapping of the PHE and TYR side chains to two site CG models. The planar nature of these side chains is critical for the spatial arrangements they occupy

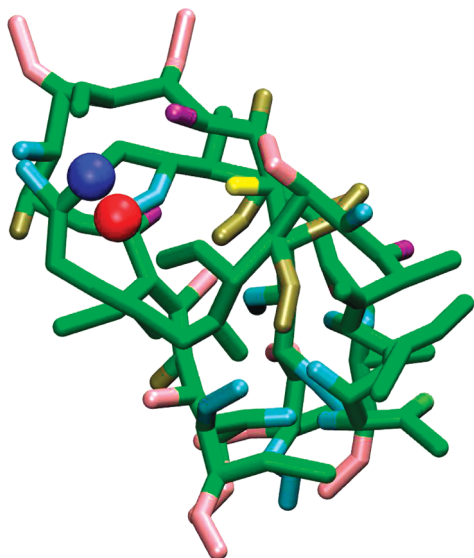


Figure 5. Shown is the CG representation of the 1dtk native structure from the lmds decoy set. The close contact is highlighted for the TYR2 site (red sphere) and the ALA site (blue sphere).

and must be conserved as much as possible in the CG model. The CG mapped structure for 1dtk native conformation is shown in Figure 5. In this figure, the TYR1 side chain site is shown as a red sphere and the ALA backbone site is shown as a blue sphere. It is easy to see here how the two site spherical representations of the phenyl based side chains leads to cases where bad contacts are generated by the CG model. This arrangement is not present in all decoys, and indeed the decoys with the lowest energy values do not have this conformation in which the TYR residue is in close proximity to other sites. The choice of a two site mapping for these residues was made to stay consistent with the mapping of roughly 3 to 4 heavy atoms. However, this issue will need to be addressed in future refinement of the model.

As mentioned above, the electrostatic residues were crudely parametrized. The current parameter set is unable to distinguish the residues ASP from GLU or LYS from ARG. This is demonstrated by the fact that the electrostatics make only small improvements to the results (*4state_reduced-1ctf* went from the number 3 ranked structure to number 1; *lmds-1igd* went from the number 2 ranked structure to number 1; others saw small or no improvement with no structures showing detriment with inclusion of electrostatics). However, it is possible that the structures studied here have conformations that are not largely dependent on electrostatic effects. Further, the percentage of charged residues is typically low in a protein, so a small impact on results is not totally unexpected.

Finally, the backbone atoms of proteins pose a particularly difficult problem as the directionality of hydrogen bonding makes the backbone an extremely anisotropic collection of atoms. This problem can be seen in the decoy sets evaluated here where high energy contacts are formed upon mapping to a CG representation. In these structures, the all atom representation shows hydrogen bonding interactions between the backbone sites leading to a stabilization of that conformation. Further, it should be mentioned that in the model

used here, the backbone was not specifically parametrized but took advantage of the CG sites developed for the ASN and GLN side chains. This is a good first approximation to the interactions of the standard backbone and ALA backbone interaction sites; however, there is room for improvement of the interaction parameters of these sites. Also, although a nonpolar spherical CG site is a crude approximation to the backbone, the results shown here suggest that the approximation can perform reasonably well. Nonetheless, improvements to the performance of the backbone CG sites could possibly begin by addition of a dipole term such as that recently employed by Cascella et al.⁷⁶

IV. Conclusions

A systematic approach has been used to develop a CG nonbonded interaction potential for amino acids. The model uses a reduced representation in which roughly 3 to 4 heavy atoms (non-hydrogen) and adjacent hydrogens are mapped to a single CG. The conservation of size in the model is reassuring for the protein–protein interaction predictive ability and for the future possibility of inclusion of a solvation contribution to the fold stability. The work here also demonstrates the consistency that can be maintained with parametrization techniques that are not based directly on protein structures. Further, it should be reiterated that no protein structural data or preconceived protein structural characteristics were used in the parametrization of this model. The encouraging results shown in the ranking of the decoy sets suggest that the current approach can be useful as a protein structure predictive tool. The decoy analysis results also highlights the importance of accurately treating the morphology of side chains especially the phenyl based residues (PHE, TRP, and TYR). As stated above, although the model is not presented here as suitable for MD simulations, there is nothing to prevent its implementation into a MD framework. Work on an intramolecular potential is currently in progress and will be presented in a forthcoming manuscript. In addition to this, further refinement of the nonbonded interactions will be a focal point of future work. The results presented herein suggest that the present approach is capable of yielding useful CG models.

Acknowledgment. The authors thank NSF, NIH, and Procter & Gamble for generous support. Also, the authors would like to thank Axel Kohlmeyer for implementation of the CG model into the LAMMPS MD code. Finally, thanks go to Vincenzo Carnevale for useful discussions. R.D. was supported by a NSF Biological Informatics Postdoctoral Fellowship under Grant No. DBI-0532800.

References

- (1) Levitt, M.; Warshel, A. *Nature* **1975**, *253*, 694–698.
- (2) Tanaka, S.; Scheraga, H. A. *Macromolecules* **1976**, *9*, 945–50.
- (3) Miyazawa, S.; Jernigan, R. L. *Macromolecules* **1985**, *18*, 534–552.
- (4) Treptow, W. L.; Barbosa, M. A. A.; Garcia, L. G.; de Araujo, A. F. P. *Proteins: Struct., Funct., Genet.* **2002**, *49*, 167–18.

- (5) Sippl, M. J. *J. Mol. Biol.* **1990**, *213*, 859–883.
- (6) Bahar, I.; Atilgan, A. R.; Erman, B. *Fold. Des.* **1997**, *2*, 173–181.
- (7) Ueda, Y.; Taketomi, H.; Go, N. *Biopolymers* **1978**, *17*, 1531–1548.
- (8) Dill, K. A.; Ozkan, S. B.; Shell, M. S.; Weikl, T. R. *Ann. Rev. Biophys.* **2008**, *37*, 289–316.
- (9) Huang, E. S.; Subbiah, S.; Levitt, M. J. *J. Mol. Biol.* **1995**, *252*, 709–720.
- (10) Schueler-Furman, O.; Wang, C.; Bradley, P.; Misura, K.; Baker, D. *Science* **2005**, *310*, 638–642.
- (11) Moulton, J. *Curr. Opin. Struct. Biol.* **2005**, *15*, 285–289.
- (12) Zhang, Y. *Curr. Opin. Struct. Biol.* **2008**, *18*, 342–348.
- (13) Shen, M. Y.; Sali, A. *Protein Sci.* **2006**, *15*, 2507–2524.
- (14) Simons, K. T.; Kooperberg, C.; Huang, E.; Baker, D. *J. Mol. Biol.* **1997**, *268*, 209–225.
- (15) Reva, B. A.; Finkelstein, A. V.; Sanner, M.; Olson, A. J.; Skolnick, J. *Protein Eng.* **1997**, *10*, 1123–1130.
- (16) Petrey, D.; Honig, B. *Mol. Cell* **2005**, *20*, 811–819.
- (17) Miyazawa, S.; Jernigan, R. L. *J. Mol. Biol.* **1996**, *256*, 623–644.
- (18) Miyazawa, S.; Jernigan, R. L. *Proteins: Struct., Funct., Genet.* **1999**, *36*, 357–369.
- (19) Sippl, M. J. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 473–501.
- (20) Sippl, M. J.; Weitckus, S. *Proteins: Struct., Funct., Genet.* **1992**, *13*, 258–271.
- (21) Kolinski, A.; Skolnick, J. *Polymer* **2004**, *45*, 511–524.
- (22) Clementi, C. *Curr. Opin. Struct. Biol.* **2008**, *18*, 10–15.
- (23) Heath, A. P.; Kaviraki, L. E.; Clementi, C. *Proteins: Struct., Funct., Bioinf.* **2007**, *68*, 646–661.
- (24) Matysiak, S.; Clementi, C. *J. Mol. Biol.* **2006**, *363*, 297–308.
- (25) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *Protein Sci.* **1993**, *2*, 1697–1714.
- (26) Liwo, A.; Oldziej, S.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 849–873.
- (27) Liwo, A.; Pincus, M. R.; Wawak, R. J.; Rackovsky, S.; Oldziej, S.; Scheraga, H. A. *J. Comput. Chem.* **1997**, *18*, 874–887.
- (28) Liwo, A.; Kazmierkiewicz, R.; Czaplowski, C.; Groth, M.; Oldziej, S.; Wawak, R. J.; Rackovsky, S.; Pincus, M. R.; Scheraga, H. A. *J. Comput. Chem.* **1998**, *19*, 259–276.
- (29) Yap, E. H.; Fawzi, N. L.; Head-Gordon, T. *Proteins: Struct., Funct., Bioinf.* **2008**, *70*, 626–638.
- (30) Khatun, J.; Khare, S. D.; Dokholyan, N. V. *J. Mol. Biol.* **2004**, *336*, 1223–1238.
- (31) Huang, E. S.; Subbiah, S.; Tsai, J.; Levitt, M. J. *J. Mol. Biol.* **1996**, *257*, 716–725.
- (32) Khurana, E.; DeVane, R.; Kohlmeyer, A.; Klein, M. L. *Nano Lett.* **2008**, *8*, 3626.
- (33) Marrink, S. J.; de Vries, A. H.; Mark, A. E. *J. Phys. Chem. B* **2004**, *108*, 750–760.
- (34) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. *J. Phys. Chem. B* **2007**, *111*, 7812–7824.
- (35) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S. J. *J. Chem. Theory Comput.* **2008**, *4*, 819–834.
- (36) Bond, P. J.; Holyoake, J.; Ivetac, A.; Khalid, S.; Sansom, M. S. P. *J. Struct. Biol.* **2007**, *157*, 593–605.
- (37) Treptow, W.; Marrink, S. J.; Tarek, M. *J. Phys. Chem. B* **2008**, *112*, 3277–3282.
- (38) Izvekov, S.; Voth, G. A. *J. Phys. Chem. B* **2005**, *109*, 2469–2473.
- (39) Noid, W. G.; Chu, J. W.; Ayton, G. S.; Voth, G. A. *J. Phys. Chem. B* **2007**, *111*, 4116–4127.
- (40) Zhou, J.; Thorpe, I. F.; Izvekov, S.; Voth, G. A. *Biophys. J.* **2007**, *92*, 4289–4303.
- (41) Tozzini, V. *Curr. Opin. Struct. Biol.* **2005**, *15*, 144–150.
- (42) Tozzini, V.; McCammon, J. A. *Chem. Phys. Lett.* **2005**, *413*, 123–128.
- (43) Masella, M.; Borgis, D.; Cuniasse, P. *J. Comput. Chem.* **2008**, *29*, 1707–1724.
- (44) Klein, M. L.; Shinoda, W. *Science* **2008**, *321*, 798–800.
- (45) Aksimentiev, A.; Schulten, K. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 4337–4338.
- (46) Arkhipov, A.; Yin, Y.; Schulten, K. *Biophys. J.* **2008**, *95*, 2806–2821.
- (47) Shih, A. Y.; Arkhipov, A.; Freddolino, P. L.; Schulten, K. *J. Phys. Chem. B* **2006**, *110*, 3674–3684.
- (48) Shelley, J. C.; Shelley, M. Y.; Reeder, R. C.; Bandyopadhyay, S.; Klein, M. L. *J. Phys. Chem. B* **2001**, *105*, 4464–4470, 26.
- (49) Shinoda, W.; DeVane, R.; Klein, M. L. *Mol. Simul.* **2006**, *33*, 27–36.
- (50) Shinoda, W.; DeVane, R.; Klein, M. L. *Soft Matter* **2008**, *4*, 2454–2462.
- (51) Bhargava, B. L.; DeVane, R.; Klein, M. L.; Balasubramanian, S. *Soft Matter* **2007**, *3*, 1395–1400.
- (52) Nielsen, S. O.; Lopez, C. F.; Srinivas, G.; Klein, M. L. *J. Chem. Phys.* **2003**, *119*, 7043–7049.
- (53) Shih, A. Y.; Freddolino, P. L.; Arkhipov, A.; Schulten, K. *J. Struct. Biol.* **2007**, *157*, 579–592.
- (54) Basdevant, N.; Borgis, D.; Ha-Duong, T. *J. Phys. Chem. B* **2007**, *111*, 9390–9399.
- (55) Han, W.; Wan, C. K.; Wu, Y. D. *J. Chem. Theory Comput.* **2008**, *4*, 1891–1901.
- (56) Plimpton, S. *J. Comput. Phys.* **1995**, *117*, 1–19.
- (57) Tuckerman, M. E.; Berne, B. J.; Martyna, G. J. *J. Chem. Phys.* **1991**, *94*, 6811–6815.
- (58) Hockney, R.; Eastwood, J. *Computer Simulation Using Particles*; IOP: Bristol, 1988.
- (59) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092.
- (60) Deserno, M.; Holm, C. *J. Chem. Phys.* **1998**, *109*, 7678–7693.
- (61) Deserno, M.; Holm, C. *J. Chem. Phys.* **1998**, *109*, 7694–7701.
- (62) Allen, M.; Tildesley, D. *Computer Simulation of Liquids*; Oxford Science: 1987.
- (63) Lee, B.; Richards, F. M. *J. Mol. Biol.* **1971**, *55*, 379.

- (64) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33.
- (65) L., Y. C. *Chemical Properties Handbook*; McGraw-Hill: 1999.
- (66) Samudrala, R.; Levitt, M. *Protein Sci.* **2000**, *9*, 1399–1401.
- (67) Simons, K. T.; Ruczinski, I.; Kooperberg, C.; Fox, B. A.; Bystroff, C.; Baker, D. *Proteins: Struct., Funct., Genet.* **1999**, *34*, 82–95.
- (68) Keasar, C.; Levitt, M. *J. Mol. Biol.* **2003**, *329*, 159–174.
- (69) Xia, Y.; Huang, E. S.; Levitt, M.; Samudrala, R. *J. Mol. Biol.* **2000**, *300*, 171–185.
- (70) Melo, F.; Sanchez, R.; Sali, A. *Protein Sci.* **2002**, *11*, 430–448.
- (71) Zhang, C.; Liu, S.; Zhou, H. Y.; Zhou, Y. Q. *Protein Sci.* **2004**, *13*, 400–411.
- (72) Zhang, C.; Liu, S.; Zhou, H. Y.; Zhou, Y. Q. *Biophys. J.* **2004**, *86*, 3349–3358.
- (73) Zhou, H. Y.; Zhou, Y. Q. *Protein Sci.* **2002**, *11*, 2714–2726.
- (74) Zhou, H. Y.; Zhou, Y. Q. *Protein Sci.* **2003**, *12*, 2121–2121.
- (75) Melo, F.; Feytmans, E. *J. Mol. Biol.* **1997**, *267*, 207–222.
- (76) Cascella, M.; Neri, M. A.; Carloni, P.; Dal Peraro, M. *J. Chem. Theory Comput.* **2008**, *4*, 1378–1385.

CT800441U

Effect of Force Field Parameters on Sodium and Potassium Ion Binding to Dipalmitoyl Phosphatidylcholine Bilayers

Arnau Cordoní,[†] Olle Edholm,[‡] and Juan J. Perez^{*,†}

Department d'Enginyeria Química, Technical University of Catalonia (UPC), Avenue Diagonal 647, 08028 Barcelona, Spain, and, Theoretical Biological Physics, Royal Institute of Technology (KTH), SE-10691 Stockholm, Sweden

Received February 12, 2009

Abstract: The behavior of electrolytes in molecular dynamics simulations of zwitterionic phospholipid bilayers is very sensitive to the force field parameters used. Here, several 200 ns molecular dynamics simulations of dipalmitoyl phosphatidylcholine (PC) bilayers in 0.2 M sodium or potassium chloride using various common force field parameters for the cations are presented. All employed parameter sets give a larger number of Na⁺ ions than K⁺ ions that bind to the lipid heads, but depending on the parameter choice quite different results are seen. A wide range of coordination numbers for the Na⁺ and K⁺ ions is also observed. These findings have been analyzed and compared to published experimental data. Some simulations produce aggregates of potassium chloride, indicating (in accordance with published simulations) that these force fields do not reproduce the delicate balance between salt and solvated ions. The differences between the force fields can be characterized by one single parameter, the electrostatic radius of the ion, which is correlated to σ_{MO} (M represents Na⁺/K⁺), the Lennard-Jones radius. When this parameter exceeds a certain threshold, binding to the lipid heads is no longer observed. One would, however, need more accurate experimental data to judge or rank the different force fields precisely. Still, reasons for the poor performance of some of the parameter sets are clearly demonstrated, and a quality control procedure is provided.

Introduction

Biological membranes are essential for cell integrity, providing a physical barrier between the interior and exterior of a cell, as well as a separation between different inner compartments. In addition, membranes provide a matrix to hold many types of proteins involved in important cell functions and, consequently, play a key role for regulation of biological processes. At physiological conditions, the membranes are in contact with an aqueous solution that contains ions like Na⁺, K⁺, Ca²⁺, Mg²⁺, and Cl⁻ and, therefore, the study of interactions between lipids, water molecules, and ions is of great importance.¹ Experiments from the last decade show

that ions play an essential role not only in modulating the structure, dynamics, and stability of membranes, but also in the binding and insertion of proteins, membrane fusion, and transport across them.² Current knowledge of the interplay between ions and biological membranes is still scarce due to experimental difficulties. However, spectroscopic experiments reveal that cations bind to negatively charged moieties in the lipid head groups of charged^{3–7} or zwitterionic phosphatidylcholine (PC) lipids,^{8–15} whereas anions, like chloride, remain solvated outside the bilayer. Moreover, images of averaged positions of ions in gel-phase PC bilayers have recently been resolved using atomic force microscopy, revealing that the negatively charged phosphate groups share the charge distributions generated by mobile cations.¹⁶

Despite early experimental evidence that ions may alter the structural properties of lipid bilayers,^{17,18} the difficulties associated with the precise characterization of these still leave

* Corresponding author. Phone: +34934016679. Fax: +34934017150. E-mail: juan.jesus.perez@upc.edu.

[†] Technical University of Catalonia.

[‡] Royal Institute of Technology.

several issues to be uncovered, especially for zwitterionic bilayers. Thus, it is well established that bound ions alter the orientation of lipid head groups,^{19–23} and there is also growing evidence that they affect the order of the lipid chains, leading to bilayer thickening. Effects on the gel to liquid crystalline phase transition of phosphatidylcholine (PC) bilayers in contact with certain mono- and divalent ions have also been reported.^{24,25} Additionally, effects on heat capacity²⁶ and nanomechanical properties of zwitterionic PC lipids²⁷ suggest that the binding of ions reduces the area per lipid. This has recently been observed in X-ray studies²⁸ of PC bilayers in solution containing Na⁺ or Ca²⁺ ions at >1 M concentrations, while an effect was not observed with K⁺ ions in the same concentration range.²⁹

Atomistic molecular dynamics (MD) simulations of lipid bilayers have emerged in the last years as a useful technique capable of providing a detailed microscopic picture of the interactions and processes occurring in biological membranes which are not always accessible experimentally. The complexity of real membranes containing different proteins, lipids, and other molecules makes atomic scale modeling of such systems computationally unaffordable. The study of one-component model lipid bilayers, however, presently is feasible and has been the focus of a large number of reports published in the last few years, being pivotal in providing new insights into these matters.^{30–36} MD simulations of lipid bilayers in contact with electrolytes support that cations bind to the head groups of zwitterionic lipids.^{5,6,26,37–43} It has, however, been observed that the equilibrium properties of such systems are sensitive to the force field parameters used for phospholipids,⁴⁴ water molecules,^{45–48} and ions⁴⁹ used in the simulations. The ion parameters are the most critical ones due to the fragmentary experimental knowledge of the structural properties of ionic solutions at finite concentration. A recent systematic comparison using different widely used parameters in simulations of an aqueous solution of sodium chloride showed that, even if all the parameters could describe the physical properties governed by the ion hydration shell in its immediate neighborhood, ion–ion radial distribution functions exhibited substantial differences.⁴⁹ In the same direction, an unbalance in the parameters of some alkali cations has recently been reported to be the origin of serious potassium chloride aggregation problems that could affect the quality of the generated MD trajectories.^{50–53}

Recent MD simulations of PC lipid bilayers published in the literature reveal large differences in the degree of K⁺ ion binding to lipid oxygens – from negligible to substantial – depending on the parameters used for K⁺ and Cl[–] ions and, therefore, on the structural electrostatic properties of the associated bilayer.^{41,43,54–56} This indicates that it is an important task to further investigate the effect of ion parameters on lipid bilayer simulations in order to improve the understanding of the different binding behavior reported for K⁺ compared to that of Na⁺ and Ca²⁺.^{28,29} The present work aims at rationalizing the effect of monovalent ion parameters on the degree of binding of such ions to PC bilayers using atomistic MD simulations. For this purpose, a set of seventeen simulations of dipalmitoyl phosphatidylcholine (DPPC) bilayers (Figure 1) in contact with 0.2 M

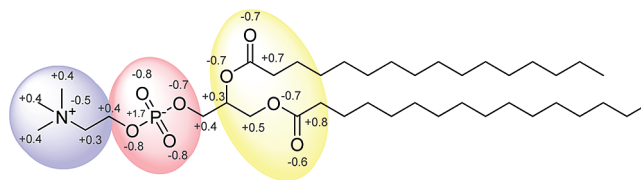


Figure 1. Chemical structure of a DPPC lipid with the partial charges given according to the Berger force field³¹ (see Methods Section). Colors indicate the following regions: choline (blue), phosphate (red), and carbonyls (yellow).

aqueous solutions of sodium or potassium chloride were carried out using an ample variety of parameters available in the literature for Na⁺ and K⁺. The present study provides a perspective on the performance of the different parameters and, furthermore, provides a guide for the selection, or further optimization, of ion parameters. Finally, it is concluded that the number of bound alkali ions per lipid drops linearly with the Lennard-Jones (LJ) parameter, σ .⁵⁷ This affects also other properties, such as the area per lipid and the head group dipole tilt angle.

Methods

Simulation Setup. All computer simulations were performed with the GROMACS 3.3.2 package^{58,59} using the same protocols described elsewhere.⁴¹ DPPC lipids were modeled using a slight modification of the parameters developed by Berger and collaborators.³¹ Parameters for alkali ions selected for the present work are listed in Table I and were retrieved from different sources. They will be referred to as: Jorgensen,⁶⁰ GROMOS,⁶¹ CHARMM,⁶² Dang,⁶³ Jensen,⁶⁴ AMBER, and OPLS (both derived from the original Åqvist parameters⁶⁵) throughout the manuscript respectively. The implementation of the OPLS-AA⁶⁶ force field in GROMACS was used for the Jorgensen and OPLS parameters. All the simulations employed Åqvist⁶⁵ parameters for Cl[–] ions and the TIP3P model for water molecules.⁶⁷

As starting configurations, a previous 40 ns simulation of a DPPC bilayer with water and sodium chloride⁴¹ was used. This system consists of 256 lipids, ~17 000 water molecules, corresponding to a hydration of 66 waters per lipid, and 61 Na⁺ and 61 Cl[–] ions, giving a concentration of sodium chloride of about 0.2 M. The initial size of the periodic simulation boxes was 8.1 × 8.1 × 12.7 nm³ (XYZ), which corresponds to an area per lipid of about 0.58 nm². The thickness of the water layer is about 8 nm, which gives a fully hydrated system by wide margins.⁴³ Systems containing KCl were generated by replacing the sodium ions in the starting box for the NaCl systems. Each simulation was run for 200 ns of which the first 100 ns were considered as equilibration and, consequently, all averages were calculated only over the last 100 ns, since it is well-known that equilibration times for this kind of system are long.⁴⁰ Figure 2 shows an equilibrated DPPC/NaCl system after 200 ns of simulations.

Details about the Ion–Lipid Potential. In a classical force field, the ion–water/lipid oxygen potential energy function consists of two terms: one electrostatic, V_{MO}^{Coul} , and

Table I. LJ Parameters for Na⁺ and K⁺ Ions Compared in the Present Study and Phosphate Lipid Oxygen Parameters

	parameter set	σ (nm)	ε (kJ/mol)	B ^g	A ^g
Na ⁺	Jorgensen ^a	0.190	6.7288	1.256×10^{-3}	5.862×10^{-8}
	CHARMM ^b	0.224	0.1963	9.998×10^{-5}	1.273×10^{-8}
	Dang ^c	0.235	0.5439	3.664×10^{-4}	6.172×10^{-8}
	GROMOS ^d	0.258	0.0618	7.207×10^{-5}	2.102×10^{-8}
	OPLS ^e	0.333	0.0116	6.335×10^{-5}	8.645×10^{-8}
	Jensen ^f	0.407	0.0021	3.804×10^{-5}	1.729×10^{-7}
K ⁺	CHARMM ^b	0.314	0.3640	1.402×10^{-3}	1.351×10^{-6}
	AMBER ^e	0.474	0.0014	6.194×10^{-5}	6.990×10^{-7}
	OPLS ^e	0.493	0.0014	7.931×10^{-5}	1.145×10^{-6}
	Jensen ^f	0.517	0.0021	1.598×10^{-4}	3.051×10^{-6}
	GROMOS ^d	0.645	0.0001	1.638×10^{-5}	1.184×10^{-6}
O	Berger	0.296	0.8780	2.362×10^{-3}	1.589×10^{-6}

^a Ref 60. ^b Ref 62. ^c Ref 63. ^d Ref 61. ^e Ref 65. ^f Ref 64. The different values for OPLS and AMBER respond to truncation on the conversions associated to each implementation. ^g B and A are given in kJ/mol·nm⁶ and kJ/mol·nm¹².

one LJ, V_{MO}^{LJ} , with M being either Na⁺ or K⁺ and O standing for oxygen:

$$V_{MO}(r) = V(r)_{MO}^{Coul} + V(r)_{MO}^{LJ}$$

Both energies are functions of the distance, r , between the atoms. There are two equivalent, widely used ways to write the LJ potential, which both will be used here:

$$V_{MO}^{LJ}(r) = 4\varepsilon_{MO} \left[\left(\frac{\sigma_{MO}}{r} \right)^{12} - \left(\frac{\sigma_{MO}}{r} \right)^6 \right] = \frac{A_{MO}}{r^{12}} - \frac{B_{MO}}{r^6}$$

In the first form, ε is the depth of the potential well, and σ is the distance at which the potential is zero, while the constants A and B , in the second expression, account for the strength of the repulsive part and the attractive parts separately. These MO parameters are obtained by averaging single atom ones, which originally have been obtained by a fit to reproduce experimental data.⁶⁸ Most force fields use a geometric mean for the ε parameters, but either a geometric or an arithmetic (also named Lorentz–Bertholet) mean may be used for the σ parameters. The geometric and arithmetic averages are similar unless parameters of the individual atoms are much different. Here simulations with both combination schemes have been performed when the different ways of averaging the σ 's have produced differences that are larger than 0.003 nm.

One of the simplest ways to model free energies of solvation of cations in water is the Born approximation:

$$\Delta G_{\text{born}} = -\frac{1}{8\pi\varepsilon_0} \left(1 - \frac{1}{\varepsilon} \right) \frac{q^2 e^2}{r_{\text{eff}}}$$

This depends only on the permittivity of the medium (ε), the charge of the cation, and the size of the cavity that it occupies (r_{eff}). The latter value is generally larger than the ionic radii (r_{ion}), since, otherwise, the expression would overestimate the energy. However, it is possible to estimate r_{eff} from the first peak in the radial distribution function (rdf) for M-water O pairs (r_{rdf})^{25,69} as

$$r_{\text{eff}} = \frac{(2r_{\text{rdf}} - r_o)}{2}$$

where, r_o is the radius of the oxygen atom in a water molecule, which is approximately 0.14 nm.

Results and Discussion

Ion Binding and Coordination. For each simulation, the average number of alkali ions bound to the DPPC bilayer was determined. It was defined as the total number of ions located closer to at least one of the lipid oxygens smaller than the distance corresponding to the first peak in the M-lipid oxygen rdf. This means within 0.31 and 0.36 nm for Na⁺ and K⁺, respectively.⁴¹ These numbers are listed as ions/lipid in Table II. Table II shows that, within every parameter set, the number of bound Na⁺ ions is always larger than that of K⁺. Moreover, between 0.09 and 0.24 Na⁺ ions bind per lipid, which is between 37 to 99% of the total number of ions. Earlier studies,⁴¹ in which higher ion concentrations also were used, indicate that sodium ions saturate the bilayer at about 0.3 Na⁺ per lipid. K⁺ ions bind to a lower extent, ranging between 0.00 and 0.14 ions per lipid, corresponding to 0 to 58% of the total number of ions. These results show that the degree of Na⁺ and K⁺ binding is extremely sensitive to force field parameters.

The ion coordination features reveal different characteristics of the binding in the simulations. In Table III, results from simulations with substantial ion binding (>0.03 ions per lipid) are listed. The coordination number (CN) of the ions, as determined from the M-water oxygen rdfs, varies between 5 and 6 for Na⁺ and between 6 and 7 for K⁺ with the phosphate and carbonyl oxygens of the lipids, water oxygens, and chloride as coordinating atoms. These numbers are in reasonable agreement with experiments as well as other computational predictions.⁷⁰ The preferred number of lipid oxygens in the first coordination shell is 4 or 5 for Na⁺, whereas for K⁺ the value range is between 2 and 5. The number of different lipids coordinating each Na⁺ is mostly 2 or 3, while this number is between 1 and 3 for K⁺. The average number of lipids per Na⁺ ion is 2.4–2.9, in good agreement with the values of 2.85 and 3.1 reported by Gurtovenko³⁹ and Baker,⁴³ despite the differences in setup and system size. We also investigated the ionic preference between the carbonyl oxygens of the glycerol backbone and phosphate oxygens with different ion parameters. All simula-

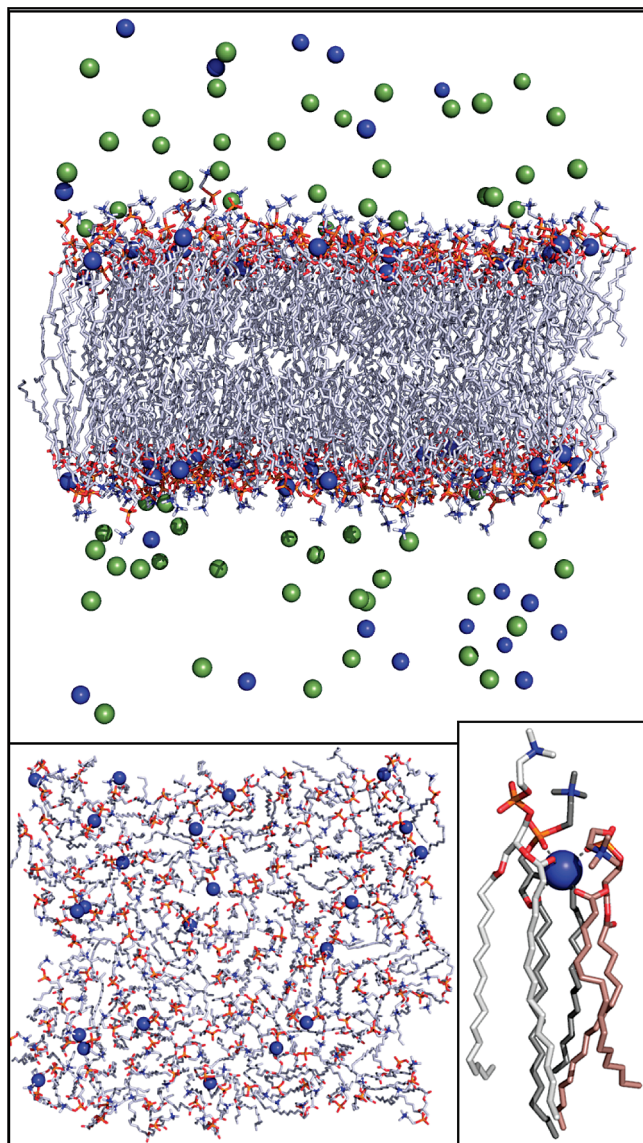


Figure 2. Snapshot of an equilibrated DPPC/NaCl system after 200 ns of MD simulations: (top) lateral view showing the bilayer and the location of Na⁺ (blue) and Cl⁻ ions (green); (bottom-left) top view of a single layer showing the bound cations; and (bottom-right) details of a representative Na⁺–DPPC complex with three lipids coordinating the ion to carbonyl and phosphate oxygens. The atoms are colored according to types, with lipid atoms represented as sticks and ions as spheres. The water molecules are not shown for better clarity.

tions suggest that both are important attractors for the ions, but that the phosphate oxygens are preferred, with average CN's of 1.9–2.7 and 1.4–2.2, for Na⁺ and K⁺, respectively, while the coordination numbers are smaller, 1.2–1.9 and 0.9–1.4, for the carbonyl oxygens (see Table III). The preference for one oxygen type or the other is not well-known experimentally,^{16,25} and it seems to be strongly dependent on the force field employed for lipids, the water model, the system size and sampling. However, the values above are useful when comparing the performance of the different cation parameters.^{38,43} It is interesting to note that both the average number of lipid oxygens coordinated to alkali ions and the number of lipids per ion bound exhibit a parallel

trend (see Table III), which correlates with the extent of ion binding shown in Table II. This suggests, as will be shown later, that differences in these features between simulations may have the same origin. They may be correlated with the strength of the cation/oxygen interaction within each parameter set, which in turn is determined by the LJ radius.

Potassium Chloride Aggregation. Before describing the structural effects associated with the ionic binding to the lipid bilayer in the next section (Structural Effects), we want to stress that some of the K⁺ parameters produce important artifacts in the simulations. Visual inspection of the final coordinates (after 200 ns) of each trajectory reveals in some cases the existence of large clusters of potassium chloride (containing essentially all ions in the system) arranged in a face-centered cubic lattice typical of alkali–chloride crystals.⁷¹ Since a salt concentration of 0.2 M is much lower than the saturation limit, 3.2 M,⁷² these aggregates should be considered as an artifact of the force field used in those simulations. This has already been observed in pure water/salt^{50,52,53,73} and water/salt/DNA^{50–52,74,75} systems. In the present work, these aggregation problems are found in the simulations using GROMOS⁶¹ and Åqvist (AMBER and OPLS) potassium parameters⁶⁸ with the Lorentz–Bertholet combination rule. This agrees with a very recent study reporting artifacts found when using the Åqvist parameters, employing the same combination rule.⁵⁰ The LJ parameters for K⁺ seem to be the main reason for this artifact, since test simulations using I⁻ or different Cl⁻ parameters did not change these results (data not shown). Second, similar aggregation problems have been reported for other water models.⁵⁰ Thus, it appears as if ϵ_{MM} values of 0.0014 kJ/mol or smaller combined with large σ_{MO} parameters leads to aggregation when the Lorentz–Bertholet combination rule is used. In contrast, the use of the geometric combination rule does not result in such artifacts. By the time this manuscript was being prepared, a corrected set of ion parameters to be used with the AMBER force field was reported in order to overcome this problem.⁵³ Interestingly, these new parameters for ion K⁺ have a much smaller σ and a larger ϵ . The use of Åqvist parameters in conjunction with geometric averaging for the σ 's in the AMBER force field has been proposed as an alternative fix.

K⁺–Cl⁻ aggregation can be monitored through the K⁺–Cl⁻ rdf at different time windows. Aggregation produces an increased second peak in the rdf at about 0.30 nm, due to transient contacts between ionic pairs, and the appearance of well-defined additional peaks at larger distances mainly at 0.55 and 0.70 nm (data not shown). The aggregation rate is slow at MD time scales, although it depends on the concentration. In the simulations presented here, typically 25–100 ns or more are required to obtain large clusters. Therefore, these artifacts were not detected until recently when such time scales became easily accessible in simulations, despite that most of the currently used ion parameters have been used for almost 20 years.^{60,68} The process is faster if the initial positions of the K⁺ ions are generated in the solvent instead of starting from sodium positions, where some are deep inside the lipid head group region (see Methods Section). A reanalysis of our previously reported 40 ns

Table II. Summary of Simulations and Results^a

cation	source	comb. rule	σ_{MO} (nm)	A_{MO}^i	area (nm ²)	dipole angle (°)	# ions per lipid	MCI aggregates	r_{eff}^j (nm)	ΔG_{hyd}^k (kJ/mol)
none	—	geom.	—	—	0.620	13	—	—	—	—
Na ⁺	Jorgensen ^b	geom.	0.237	3.05×10^{-7}	0.531	26	0.24	No	0.158	-434
	CHARMM ^c	geom.	0.260	1.60×10^{-7}	0.538	27	0.23	No	0.146	-470
	Dang ^d	geom.	0.266	3.47×10^{-7}	0.541	27	0.22	No	0.158	-434
	GROMOS ^e	geom.	0.276	1.82×10^{-7}	0.541	26	0.22	No	0.148	-463
	OPLS ^f	geom.	0.314	3.71×10^{-7}	0.562	22	0.16	No	0.166	-413
		arit.	0.315	3.78×10^{-7}	0.570	22	0.15	No	0.166	-413
	Jensen ^g	geom.	0.347	5.22×10^{-7}	0.576	19	0.11	No	0.176	-390
		arit.	0.352	6.10×10^{-7}	0.593	18	0.09	No	0.181	-380
	K ⁺	CHARMM ^c	geom.	0.305	1.47×10^{-6}	0.579	21	0.14	No	0.198
GROMOS ^e		geom.	0.437	1.37×10^{-6}	0.602	15	0.04	No	0.202	-340
		arit.	0.471	3.34×10^{-6}	0.617	13	0.00	Yes	0.226	-303
		geom.	0.382	1.35×10^{-6}	0.600	16	0.04	no	0.202	-340
OPLS ^f		arit.	0.395	1.99×10^{-6}	0.618	13	0.00	yes	0.215	-320
		geom.	0.374	1.04×10^{-6}	0.594	13	0.06	no	0.194	-353
AMBER ^f		arit.	0.385	1.47×10^{-6}	0.619	13	0.00	yes	0.206	-333
		geom.	0.391	2.19×10^{-6}	0.603	14	0.03	no	0.214	-320
Jensen ^g		geom.	0.391	2.19×10^{-6}	0.603	14	0.03	no	0.214	-320
		arit.	0.407	3.49×10^{-6}	0.615	13	0.01	no	0.226	-303

^a The combination rule accounts for how σ pairs are computed as a geometric or an arithmetic average. The standard deviations are in area per lipid 0.005–0.008 nm², in number of bound ions per lipid less than 0.01, while the PN dipole angle has a broad distribution with a width of about 35°. ^b Ref 60. ^c Ref 62. ^d Ref 63. ^e Ref 61. ^f Ref 65. ^g Ref 64. ^h A in kJ/mol. ⁱ nm¹². ^j $r_{eff} = (2r_{df} - r_0)/2$. ^k ΔG_{hyd} from the Born approximation (see Methods Section).

Table III. Average Number of Lipid Oxygens of Different Types and Average Number of Lipids Coordinated to Each Bound Na⁺ and K⁺ ion Together with Standard Deviations

cation	comb. rule	no. of phosphate oxygens	no. of carbonyl oxygens	total no. of lipid oxygens	no. of lipids	
Na ⁺	Jorgensen ^a	geom.	1.82 ± 0.07	2.65 ± 0.08	5.36 ± 0.08	2.92 ± 0.05
	CHARMM ^b	geom.	1.87 ± 0.07	2.30 ± 0.08	4.41 ± 0.08	2.78 ± 0.06
	Dang ^c	geom.	1.82 ± 0.07	2.47 ± 0.12	4.62 ± 0.12	2.82 ± 0.07
	GROMOS ^d	geom.	1.78 ± 0.07	2.31 ± 0.10	4.27 ± 0.13	2.70 ± 0.06
	OPLS ^e	geom.	1.56 ± 0.11	2.06 ± 0.11	4.07 ± 0.15	2.51 ± 0.09
arit.		1.53 ± 0.14	2.14 ± 0.17	3.94 ± 0.2	2.51 ± 0.12	
Na ⁺	Jensen ^f	geom.	1.33 ± 0.15	2.03 ± 0.20	3.61 ± 0.26	2.41 ± 0.15
		arit.	1.24 ± 0.15	1.86 ± 0.21	3.36 ± 0.27	2.21 ± 0.15
K ⁺	CHARMM ^b	geom.	1.42 ± 0.15	2.24 ± 0.19	4.40 ± 0.28	2.49 ± 0.13
	GROMOS ^d	geom.	1.02 ± 0.24	1.59 ± 0.35	2.92 ± 0.45	1.96 ± 0.26
	OPLS ^e	geom.	0.99 ± 0.24	1.61 ± 0.47	2.93 ± 0.45	1.97 ± 0.25
	AMBER ^e	geom.	1.13 ± 0.18	1.73 ± 0.29	3.26 ± 0.45	2.08 ± 0.21
	Jensen ^f	geom.	0.89 ± 0.30	1.38 ± 0.44	2.61 ± 0.55	1.82 ± 0.32

^a Ref 60. ^b Ref 62. ^c Ref 63. ^d Ref 61. ^e Ref 65. ^f Ref 64.

simulation of DPPC bilayers with potassium chloride⁴¹ reveals that aggregation had begun, although not in the form of a large single cluster as reported here. Although it is obvious that the existence of aggregates limits the amount of free K⁺ ions that may interact with the lipid oxygens, our results still suggest that the difference between small or negligible binding does not necessarily come from such artifacts. Finally, it is clear that the presence of these aggregates in the vicinity of the lipid bilayer does not affect the bilayer.

Structural Effects. There is ample discussion in the literature about how binding of different ions affects the ordering of the acyl chains as well as the orientation of the lipid head dipole in simulations. This includes alkaline and alkaline–earth ones.^{26,40–42,44} Therefore, the present analysis will focus on differences due to force field parameters for the same ion. Lipid ordering may be monitored in different ways, but it is most easily observed as a decrease in area per lipid. The average values from the different simulations are shown in Table II. The simulation without

ions exhibits an area per lipid of 0.620 nm², close to the experimental value of 0.64 nm². For the systems containing 0.2 M MCI (M represents K⁺ or Na⁺), the area is in some cases considerably smaller, in agreement with the previously reported results.⁴¹ In the simulations containing Na⁺ ions, the area per lipid varies from 0.531 to 0.593 nm², while those with K⁺ ions show areas from 0.579 nm² up to the value in the salt-free system. In summary, the present simulations show that the binding of ions may reduce the area per lipid up to 15%. The data in Table II also shows a strong correlation between the decrease in area per lipid and the number of bound ions, which is clearly illustrated when these variables are plotted against each other in Figure 3.

The charge distribution in the PC head group can be represented by a dipole moment that is mainly due to the negatively charged phosphate group and the positively charged choline group. The contribution from each of the carbonyl groups is about 1 order of magnitude smaller. The orientation of the head group is sensitive to the local electrostatic field generated by neighboring atoms. Therefore,

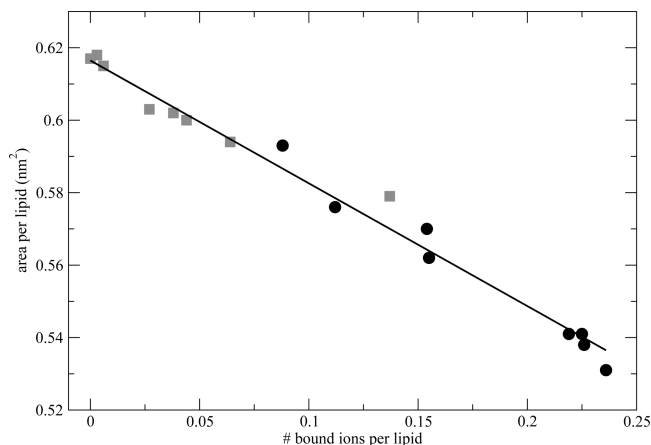


Figure 3. Area per lipid versus number of K^+ (gray squares) or Na^+ (black spheres) ions bound to the lipid bilayer. The line is a linear regression to both types of ions.

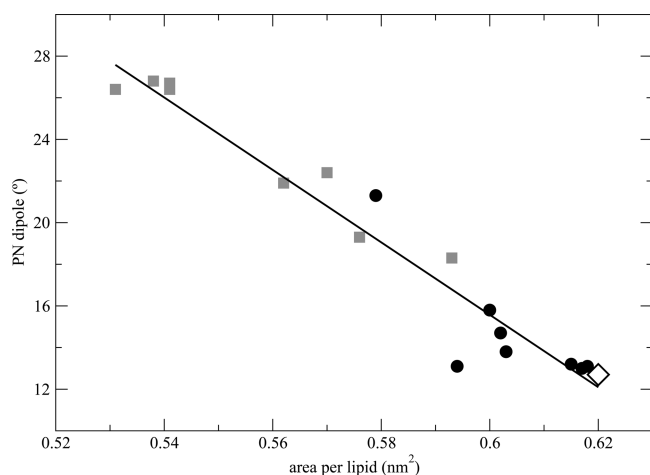


Figure 4. The **P–N** dipole tilt out of the membrane plane versus area per lipid. Data are displayed for Na^+ as black spheres and K^+ as gray squares. The line is a linear regression to both ions types. The unfilled diamond indicates the system without salt.

the orientation of the head group dipole is sensitive to the presence of ions^{26,37,38,41} and acts as a charge sensor.^{22,23} Here, the orientation of the dipole is measured as the (time and system) averaged angle between the **P–N** vectors and the bilayer plane. The results displayed in Table II show that whereas the system without ions exhibits an average angle of 13° , the binding of ions results in an increase of this value by $7\text{--}13^\circ$. Interestingly, the changes in the **P–N** tilt and the area per lipid are correlated as shown in Figure 4. This effect has already been observed when comparing simulations performed using different cutoff methods.⁷⁶ The reason for this correlation appears to be the balance between electrostatic energy and lipid packing. Thus, in the absence of ions, the **P–N** dipole tends to lie down close to the membrane plane reducing the electrostatic energy of the system. When ions are present, they screen the electrostatic interactions between the head groups and diminish the need of having the dipole close to the membrane plane. The tilt out of the membrane plane favors a closer lipid packing and reduces the area per lipid. In a similar way Grutovenko³⁹ correlated the presence of a chloride ion close a bilayer (within 1 nm

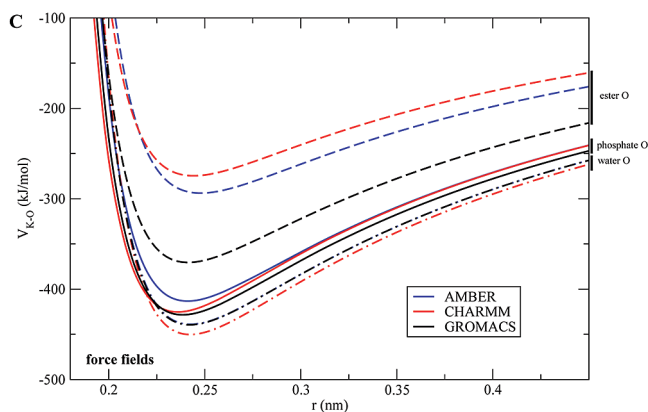
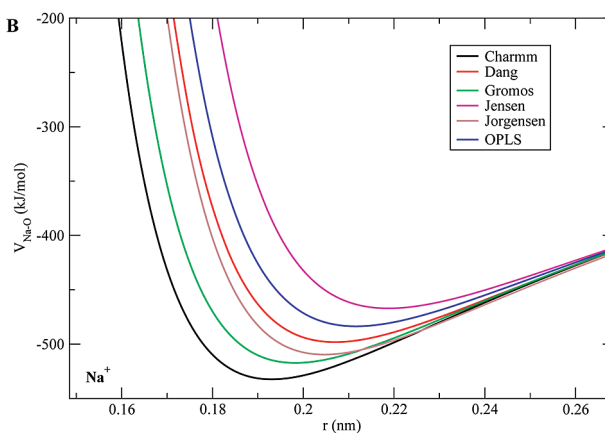
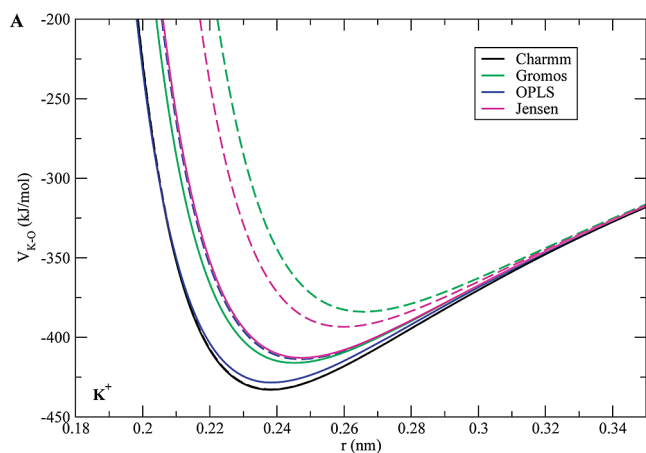


Figure 5. Cation–water/lipid oxygen potential energy function (Coulomb + LJ terms) for: (A) K^+ –lipid phosphate oxygen pairs, (B) Na^+ –lipid phosphate oxygen pairs, and (C) various K^+ –water/lipid oxygen functions in different force fields (AMBER, CHARMM, and GROMOS, see Table IV). Full lines represent geometric and dashed lines arithmetic combination rules for σ . Results in B are only shown using geometric averages of σ 's since the arithmetic averages produce results that are very similar (see Table II). Combination rules for σ are arithmetic in AMBER and CHARMM force fields and geometric for Berger's. Oxygen parameters are shown in Table IV. Corresponding AMBER, CHARMM, and OPLS K^+ parameters are displayed in Table I.

of a choline) with an increased dipole tilt of 12° out of the membrane plane.

The σ_{MO} Parameter. We also observe that the change in area per lipid and orientation of the lipid head group varies

Table IV. Fractional Charges and LJ Parameters for Lipid Phosphate, Lipid Ester, and Water Oxygens in Three Common Lipid Force Fields

Force field	atom	parameter	charge (<i>e</i>)	σ (nm)	ϵ (kJ/mol)
AMBER/GAFF	phosphate O	o	-0.780	0.2960	0.8786
	ester O	o	-0.570	0.2960	0.8786
	water O	ow (TIP3P)	-0.834	0.3151	0.6368
CHARMM	phosphate O	O2L	-0.780	0.3029	0.5021
	ester O	OBL	-0.520	0.3029	0.5021
	water O	OW (SPC)	-0.848	0.3166	0.6502
BERGER/OPLS	phosphate O	LO2	-0.800	0.2960	0.8780
	ester O	LO2	-0.700 ^a	0.2960	0.8780
	water O	opls_111 (TIP3P)	-0.834	0.3151	0.6368

^a This charge is -0.6 in one of the two chains (closer to those in the remaining force fields compared).

in a similar way as the ion hydration energies between the force fields (see Methods Section). Therefore, the degree of ion binding to the lipid bilayer is related to the energy of hydration that results from the parameters used, for which an estimate is given in Table II using the Born formula (see Methods Section). The average hydration energies are -425 ± 24 kJ/mol for Na^+ and -329 ± 17 kJ/mol for K^+ , relatively close to the values of -382 and -302 kJ/mol obtained from experiments⁷⁷ and in good agreement with recent extensive simulations.⁷⁸ The main difference between the force fields studied here is the cation–oxygen potential energy function. In order to illustrate this difference, the K^+ –O and Na^+ –O interaction energy functions are shown in Figure 5A and B for different metal ion parameters but the same oxygen parameters (those of the phosphate oxygens of the lipid model used here). The position of the K^+ –O minimum varies from 0.23 to 0.36 nm. In some cases, there are also substantial differences depending on whether the LJ radius was calculated as a geometric or arithmetic average. For the Na^+ ions, the differences are much smaller but still lead to distinctly different locations of the minima ranging from 0.19 to 0.22 nm. In the same way as in Figures 5A and B, we show in Figure 5C the K^+ –water/lipid oxygen potential energy functions for different types of oxygens in a few widely used force fields (see Table IV). It should be noted that the energy and location of the minima are similar for the different functions involving water oxygens and also lipid phosphate oxygens for the different water models and the K^+ parameters. The reasons for this are that the oxygen charge is similar in these cases (about -0.8), the cation parameters are transferable between SPC and TIP3P⁶⁵ water models and also the parameters compared provide similar Born energies. On the contrary, the function involving ester oxygens differs much more, mainly since the ester oxygen charges differ between the force fields (see Table IV). The larger carbonyl dipole in the Berger force field could explain the difference in preference of cations for phosphate or carbonyl lipid oxygens between simulations performed with this and other lipid force fields.⁴³

The electrostatic attraction dominates the cation–oxygen interaction at long distances, while LJ repulsion dominates at short distances. The attractive part of the LJ interaction is, thus, always much smaller and, in most cases, negligible compared to the electrostatic contribution to the potassium–oxygen interaction. Since the electrostatic part is the same

in all cases, it is the strength of the repulsive part of the LJ potential that controls the distance between the cation and the oxygen and the strength of binding. This is controlled by A_{MO} or σ_{MO} parameters, depending on which form of the LJ equation is used (see Table II and Methods Section).

Figure 6A shows the number of ions bound to the lipid oxygens versus σ_{MO} in the different simulations. There is a linear correlation between these two variables for σ_{MO} below 0.382 nm, while ion binding is weak or nonexistent for larger values. Similarly, Figures 4B and C show that both the area per lipid and the orientation of the lipid head group exhibit a similar linear correlation with σ_{MO} below 0.382 nm. The use of $A_{\text{MO}}^{1/2}$ instead of σ_{MO} as an independent variable gives similar linear correlations (not shown). The main result can be summarized as there is a threshold for σ_{MO} (or $A_{\text{MO}}^{1/2}$) above which ion binding is negligible. Above this value, the change in area per lipid or lipid head group dipole angle is, therefore, negligible. Below the threshold ion binding, areas per lipid as well as head group dipole tilt change linearly with the distance from the threshold.

Conclusions

The present work constitutes a systematic study of the effect of sodium and potassium LJ parameters on the ion binding to DPPC bilayers. For this purpose, a set of 200 ns simulations of DPPC bilayers in contact with 0.2 M aqueous salt solutions using different cation LJ parameters were performed. Eight trajectories with sodium chloride, nine with potassium chloride, and one without salt were generated.

The K^+ ions bind less to PC bilayers than Na^+ ions due to their larger radius. There are, however, large differences in binding of both ions depending on LJ parameters. For 0.2 M NaCl, different parameters give between 0.09 and 0.24 bound Na^+ ions per lipid, while the number of bound K^+ ions per lipid varies between 0 and 0.14 at the same concentration. The reason for the difference is that a larger ion has difficulties entering the head group region of the lipid bilayer. The size of the ions is determined by the balance between repulsive and attractive forces.

If these are just LJ ones, the σ_{MO} LJ parameter (where M is the cation and O is an oxygen) would be a good measure of the size. The LJ attraction between oppositely charged atoms is, however, negligible compared to that

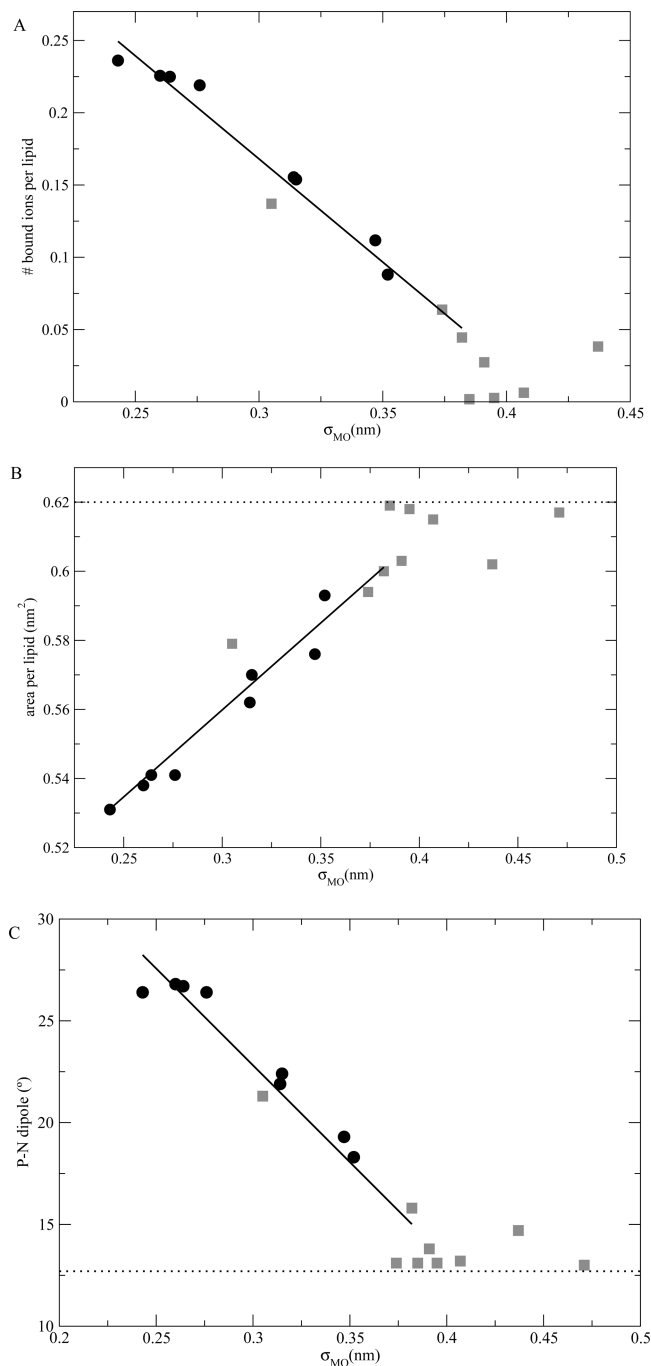


Figure 6. (A) Number of bound ions per lipid, (B) area per lipid, and (C) lipid head group (P–N) dipole tilt angle out the membrane plane versus the σ_{MO} parameters. The results are displayed for Na^+ as black spheres and K^+ as gray squares. The full lines are linear regressions excluding the simulations with $\sigma_{MO} > 0.382$ nm. The dotted lines in B and C indicate the area and the dipole tilt in the system without salt.

of the electrostatic attraction. Thus, one might argue that it would be more proper to use a measure for the radius that only includes the repulsive LJ parameter A . This parameter can, however, be expressed as $4\epsilon\sigma$.⁴¹ The electrostatic radius is, thus, proportional to $\sigma\epsilon$ and only very weakly dependent upon the other van der Waals parameter, ϵ . This explains why only small differences are seen depending on which of the measures are used

for the radius. The number of bound ions decreases with increasing radius, with a good linear correlation up to a maximum value of $\sigma_{MO} = 0.382$ nm above which binding disappears. This is reflected in area per lipid and helix dipole tilt, which covary with the amount of bound ions. The present results agree with our previous hypothesis that monovalent ions larger than K^+ (Rb^+ and Cs^+) do not bind to PC bilayers,⁴¹ since their σ_{MO} are even larger than that of K^+ .

Despite the differences observed with different force field parameters, the lack of accurate enough experimental results prohibits discerning the best parameters to be used in this type of simulation. However, the present results should be useful for selecting or calibrating new parameters when additional experimental information becomes available. It is clear that in systems where ions interact with large electrostatic dipoles or oppositely charged atoms, the attraction is dominated by electrostatics. The van der Waals attraction will then only play a negligible role and only the repulsive LJ (A) can be accurately determined from data for such systems. To determine the attractive parameter, B , one would need data for systems in which these LJ attraction interactions dominate, or at least are comparable to, electrostatic attraction. Such systems are rare, but this means also that the B parameter is of much less interest than the repulsive A parameter for ions. In any case, the present simulations show that the GROMOS and Åquist K^+ parameters should not be used with the Lorentz–Bertholet combination rule (arithmetic averaging of the σ 's), since this produces unrealistic potassium chloride aggregation as an artifact. Both parameters were designed to be used with the other (geometrical) combination rule. Moreover, a new set of alkali and halide ion parameters became available for AMBER to overcome such aggregation problems⁵³ during the preparation of this manuscript. The new σ for K^+ has then been reduced and would give results closer to remaining sets tested here. This constitutes a warning not to mix parameters that are not self-consistent without careful tests. It is straightforward to extend the calibration scheme proposed here for Na^+ and K^+ to other ions, as long as the interactions can be properly modeled by classical pairwise electrostatic and LJ terms. Finally, the trends presented here using the TIP3P water model should also be valid for the other three site water models with minor quantitative differences, keeping in mind the slight differences between these models.

It is still uncertain if MD simulations to some extent exaggerate⁴³ the ion binding to lipid bilayers compared to the few experimental results reported thus far.^{28,29} When comparing simulations, the degree of K^+ binding (if any) is still an open question. We do, here, reproduce the trend of previous studies in the sense that the larger the alkali ion, the less it binds to the membrane.^{9,16,24,25,37,41} Moreover, present results reconcile previous MD studies of the binding of ions to PC bilayers.^{41–43,54,56} However, there are other issues that need to be clarified, including other force field weaknesses and sampling artifacts.

Acknowledgment. The Spanish Ministry of Science and Technology supported this work through grant number SAF2008-04943-C02-01.

References

- (1) Berkowitz, M. L.; Bostick, D. L.; Pandit, S. *Chem. Rev.* **2006**, *106*, 1527.
- (2) Parsegian, V. A. *Ann. N.Y. Acad. Sci.* **1975**, *264*, 161.
- (3) Elmore, D. E. *FEBS Lett.* **2006**, *580*, 144.
- (4) Zhao, W.; Rog, T.; Gurtovenko, A. A.; Vattulainen, I.; Karttunen, M. *Biophys. J.* **2007**, *92*, 1114.
- (5) Pandit, S. A.; Berkowitz, M. L. *Biophys. J.* **2002**, *82*, 1818.
- (6) Mukhopadhyay, P.; Monticelli, L.; Tieleman, D. P. *Biophys. J.* **2004**, *86*, 1601.
- (7) Pedersen, U. R.; Leidy, C.; Westh, P.; Peters, G. H. *Biochim. Biophys. Acta* **2006**, *1758*, 573.
- (8) Loosley-Millman, M. E.; Rand, R. P.; Parsegian, V. A. *Biophys. J.* **1982**, *40*, 221.
- (9) Tatulian, S. A. *Eur. J. Biochem.* **1987**, *170*, 413.
- (10) Cunningham, B. A.; Gelerinter, E.; Lis, L. J. *J. Chem. Phys. Lipids* **1988**, *46*, 205.
- (11) Altenbach, C.; Seelig, J. *Biochemistry* **1984**, *23*, 3913.
- (12) Lis, L. J.; Lis, W. T.; Parsegian, V. A.; Rand, R. P. *Biochemistry* **1981**, *20*, 1771.
- (13) Lis, L. J.; Parsegian, V. A.; Rand, R. P. *Biochemistry* **1981**, *20*, 1761.
- (14) Herbert, L.; Napolitano, C. A.; McDaniel, R. V. *Biophys. J.* **1984**, *46*, 677.
- (15) Akutsu, H.; Seelig, J. *Biochemistry* **1981**, *20*, 7366.
- (16) Fukuma, T.; Higgins, M. J.; Jarvis, S. P. *Phys. Rev. Lett.* **2007**, *98*, 106101.
- (17) Watts, A.; Harlos, K.; Marsh, D. *Biochim. Biophys. Acta* **1981**, *645*, 91.
- (18) Hauser, H.; Shipley, G. G. *Biochemistry* **1984**, *23*, 34.
- (19) Roux, M.; Bloom, M. *Biochemistry* **1990**, *29*, 7077.
- (20) Clarke, R. J.; Lupfert, C. *Biophys. J.* **1999**, *76*, 2614.
- (21) Brown, M. F.; Seelig, J. *Nature* **1977**, *269*, 721.
- (22) Akutsu, H.; Nagamori, T. *Biochemistry* **1991**, *30*, 4510.
- (23) Seelig, J.; Macdonald, P. M.; Scherer, P. G. *Biochem.* **1987**, *26*, 7535.
- (24) Chapman, D.; Peel, W. E.; Kingston, B.; Lilley, T. H. *Biochim. Biophys. Acta* **1977**, *464*, 260.
- (25) Binder, H.; Zschornig, O. *Chem. Phys. Lipids* **2002**, *115*, 39.
- (26) Böckmann, R. A.; Hac, A.; Heimburg, T.; Grubmüller, H. *Biophys. J.* **2003**, *85*, 1647.
- (27) Garcia-Manyes, S.; Oncins, G.; Sanz, F. *Biophys. J.* **2005**, *89*, 1812.
- (28) Pabst, G.; Hodzic, A.; Strancar, J.; Danner, S.; Rappolt, M.; Laggner, P. *Biophys. J.* **2007**, *93*, 2688.
- (29) Petrache, H. I.; Tristram-Nagle, S.; Harries, D.; Kucerka, N.; Nagle, J. F.; Parsegian, V. A. *J. Lipid Res.* **2006**, *47*, 302.
- (30) van der Ploeg, P.; Berendsen, H. J. C. *J. Chem. Phys.* **1982**, *76*, 3271.
- (31) Berger, O.; Edholm, O.; Jähnig, F. *Biophys. J.* **1997**, *72*, 2002.
- (32) Feller, S. E. *Curr. Opin. Struct. Biol.* **2000**, *5*, 217.
- (33) Tieleman, D. P.; Marrink, S. J.; Berendsen, H. J. C. *Biochim. Biophys. Acta* **1997**, *1331*, 235.
- (34) Smondyrev, A. M.; Berkowitz, M. L. *J. Comput. Chem.* **1999**, *20*, 531.
- (35) Lindahl, E.; Edholm, O. *Biophys. J.* **2000**, *79*, 426.
- (36) Feller, S. E.; Gawrisch, K.; MacKerell, A. D., Jr. *J. Am. Chem. Soc.* **2002**, *124*, 318.
- (37) Pandit, S. A.; Bostick, D.; Berkowitz, M. L. *Biophys. J.* **2003**, *84*, 3743.
- (38) Sachs, J. N.; Nanda, H.; Petrache, H. I.; Woolf, T. B. *Biophys. J.* **2004**, *86*, 3772.
- (39) Gurtovenko, A. A. *J. Chem. Phys.* **2005**, *122*, 244902.
- (40) Böckmann, R. A.; Grubmüller, H. *Angew. Chem., Int. Ed.* **2004**, *43*, 1021.
- (41) Cordoní, A.; Edholm, O.; Perez, J. J. *J. Phys. Chem. B* **2008**, *112*, 1397.
- (42) Gurtovenko, A. A.; Vattulainen, I. *J. Phys. Chem. B* **2008**, *112*, 1953.
- (43) Lee, S. J.; Song, Y.; Baker, N. A. *Biophys. J.* **2008**, *94*, 3565.
- (44) Siu, S. W.; Vacha, R.; Jungwirth, P.; Bockmann, R. A. *J. Chem. Phys.* **2008**, *128*, 125103.
- (45) Mark, P.; Nilsson, L. *J. Comput. Chem.* **2002**, *23*, 1211.
- (46) Wallqvist, A.; Mountain, R. D. *Rev. Comput. Chem.* **1999**, *13*, 183.
- (47) Guillot, B. *J. Mol. Liq.* **2002**, *101*, 219.
- (48) Jorgensen, W. L.; Tirado-Rives, J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6665.
- (49) Patra, M.; Karttunen, M. *J. Comput. Chem.* **2004**, *25*, 678.
- (50) Auffinger, P.; Cheatham, T. E.; Vaiana, A. C. *J. Chem. Theory Comput.* **2007**, *3*, 1851.
- (51) Chen, A. A.; Pappu, R. V. *J. Phys. Chem. B* **2007**, *111*, 11884.
- (52) Chen, A. A.; Pappu, R. V. *J. Phys. Chem. B* **2007**, *111*, 6469.
- (53) Joung, I. S.; Cheatham, T. E. *J. Phys. Chem. B* **2008**, *112*, 9020.
- (54) Gurtovenko, A. A.; Vattulainen, I. *Biophys. J.* **2007**, *92*, 1878.
- (55) Gurtovenko, A. A.; Vattulainen, I. *J. Phys. Chem. B* **2008**, *112*, 1953.
- (56) Lee, S. J.; Song, Y.; Baker, N. A. *Biophys. J.* **2008**, *94*, 3565.
- (57) Lennard-Jones, J. E. *Proc. Phys. Soc., London* **1931**, *43*, 461.
- (58) Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. *Comput. Phys. Commun.* **1995**, *91*, 43.
- (59) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Model.* **2001**, *7*, 306.
- (60) Chandrasekhar, J.; Spellmeyer, D. C.; Jorgensen, W. L. *J. Am. Chem. Soc.* **1984**, *106*, 903.
- (61) Straatsma, T. P.; Berendsen, H. J. C. *J. Chem. Phys.* **1988**, *89*, 5876.
- (62) Beglov, D.; Roux, B. *J. Chem. Phys.* **1994**, *100*, 9050.

- (63) Dang, L. X.; Kollman, P. A. *J. Phys. Chem.* **1995**, *99*, 55.
- (64) Jensen, K. P.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2006**, *2*, 1499.
- (65) Åqvist, J. *J. Phys. Chem.* **1994**, *98*, 8253.
- (66) Jorgensen, W. L.; Maxwell, D. S.; TiradoRives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225.
- (67) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
- (68) Åqvist, J. *J. Phys. Chem.* **1990**, *94*, 8021.
- (69) Babu, C. S.; Lim, C. *Chem. Phys. Lett.* **1999**, *310*, 225.
- (70) Tunell, I.; Lim, C. *Inorg. Chem.* **2006**, *45*, 4811.
- (71) Degreve, L.; Borin, A. C.; Mazze, F. M.; Rodrigues, A. L. G. *Chem. Phys.* **2001**, *265*, 193.
- (72) Stephen, H.; Stephen, T.; *Solubilities of Inorganic and Organic Compounds*; Pergamon Press: Oxford, U.K., 1963.
- (73) Vieira, D. S.; Degreve, L. *J. Mol. Struct. - THEOCHEM* **2002**, *580*, 127.
- (74) Mazur, A. K. *J. Am. Chem. Soc.* **2003**, *125*, 7849.
- (75) Cheatham, T. E. *Curr. Opin. Struct. Biol.* **2004**, *14*, 360.
- (76) Wohlert, J.; Edholm, O. *Biophys. J.* **2004**, *87*, 2433.
- (77) Marcus, Y. *Ion Properties*; Dekker: New York 1997.
- (78) Kastenholz, M. A.; Hunenberger, P. H. *J. Chem. Phys.* **2006**, *124*.

CT9000763

Determination of Free Energy Profiles for the Translocation of Polynucleotides through α -Hemolysin Nanopores using Non-Equilibrium Molecular Dynamics Simulations

Hugh S. C. Martin,[†] Shantenu Jha,^{*,‡} Stefan Howorka,[†] and Peter V. Coveney^{*,†}

Centre for Computational Science and Department of Chemistry, University College London, 20 Gordon Street, London, United Kingdom, and Center for Computation and Technology, Louisiana State University, Baton Rouge, Louisiana 70803

Received February 21, 2009

Abstract: The translocation of polynucleotides through transmembrane protein pores is a fundamental biological process with important technological and medical relevance. The translocation process is complex, and it is influenced by a range of factors including the diameter and inner surface of the pore, the secondary structure of the polymer, and the interactions between the polymer and protein. In this paper, we perform nonequilibrium constant velocity-steered molecular dynamics simulations of nucleic acid molecule translocation through the protein nanopore α -hemolysin and use Jarzynski's identity to determine the associated free energy profiles. With this approach we are able to explain the observed differences in experimental translocation time through the nanopore between polyadenosine and polydeoxycytidine. The translocation of polynucleotides and single nucleotides through α -hemolysin is investigated. These simulations are computationally intensive as they employ models with atomistic level resolution; in addition to their size, these systems are challenging to study due to the time scales of translocation of large asymmetric molecules. Our simulations provide insight into the role of the interactions between the nucleic acid molecules and the protein pore. Mutated protein pores provide confirmation of residue-specific interactions between nucleotides and the protein pore. By harnessing such molecular dynamics simulations, we gain new physicochemical insight into the translocation process.

1. Introduction

The translocation of nucleic acid strands through confined protein pores has biological relevance in instances such as transfer of antibiotic resistance genes between bacteria,^{1–3} phage infection,⁴ and the uptake of oligonucleotides into kidney tissue.⁵ Moreover, the passage of nucleic acids through pores is also of biotechnological and diagnostic relevance. For these applications, a single nanopore is inserted into a lipid bilayer, and individual negatively charged

nucleic acids are electrophoretically driven through the pore. The passages of strands lead to detectable fluctuations in the ionic pore current. Data from these single channel current recording (SCCR) experiments provide information on polymer length, orientation, and composition for polymers such as single stranded DNA and RNA.^{6–11} The capability of SCCR to reveal information on translocating DNA strands may provide an avenue for faster and cheaper genetic sequencing. Progress toward single base detection has been realized; for example, a modified nanopore has been constructed to yield a base-specific current response from isolated ribonucleoside and deoxyribonucleoside monophosphates,¹² while modified SCCR techniques provide measurable current responses to single mutations in a static nucleic

* Corresponding authors. E-mail: sjha@cct.lsu.edu (S.J.) and p.v.coveney@ucl.ac.uk (P.V.C.).

[†] University College London.

[‡] Louisiana State University.

acid polymer.^{13,14} Recently, oligonucleotides have been covalently modified with chemical tags as a step towards base-specific sensing in nanopore current recordings.¹⁵

Understanding the microscopic processes of nucleic acid translocation through nanopores is crucial in improving SCCR techniques and apparatus for sequencing DNA. Using molecular dynamics simulations of the translocation process, it is in general possible to retrieve kinetic and structural information that cannot be obtained solely through experiments. Experiments investigating the translocation of nucleic acid under the influence of a transmembrane potential suggest that the process typically takes hundreds of microseconds to tens of milliseconds.⁷ But accurately simulating biological processes and systems with atomistic resolution remains a challenge for many reasons, not least of which are the very substantial computational resources required. Even with state-of-the-art high-end computers, performing simulations with atomistic resolution for such large systems over the required time scales remains currently infeasible. Applying an artificially high potential difference in the simulations to induce faster translocation disrupts the lipid membrane; applying a high uniform electrostatic field to only the translocating atoms fails to translocate nucleic acid polymers through the protein nanopore.¹⁶ Thus, if these events are to be effectively investigated, novel approaches and better algorithms are required in order to bridge the gap between time scales over which the translocation events occur and those which are accessible using simple equilibrium simulations.

Nonequilibrium simulations offer one possible approach to increasing the time scales over which physical phenomena may be investigated with atomistic resolution. Various types of nonequilibrium simulations can be designed to effect translocation at higher speeds, allowing the full translocation process to be studied. Steered molecular dynamics¹⁷ is a type of nonequilibrium simulation that applies either a constant force to an atom or group of atoms or connects an atom or center of mass of a group of atoms via a harmonic spring (governed by a force or spring constant, k) to a constraint position, which is moved at a constant velocity. Constant velocity-steered molecular dynamics (cv-SMD) has the advantage of a well-defined wall clock and simulated time frame for a given translocation distance; cv-SMD is used in this paper to induce high-speed translocation.

This paper exploits nonequilibrium molecular dynamics simulations to gain microscopic insight into nucleic acid translocation through protein pores from the calculation of free energy profiles. It is desirable to obtain the free energy profile of the translocation process due to its central role in determining the thermodynamic properties of a system; it can also be used to ascertain kinetic properties. Free energy is an equilibrium property; in order to compute the free energy from a nonequilibrium process such as cv-SMD, we use Jarzynski's equality (JE).¹⁸ By obtaining and analyzing the free energy profiles of the translocation process, information is revealed about the energy barriers that a translocating nucleotide molecule experiences, and physical insight into the nature and origins of these barriers is gained.

In this paper, the protein nanopore we use is α -hemolysin (α HL), depicted in Figure 1. α HL is a heptameric protein

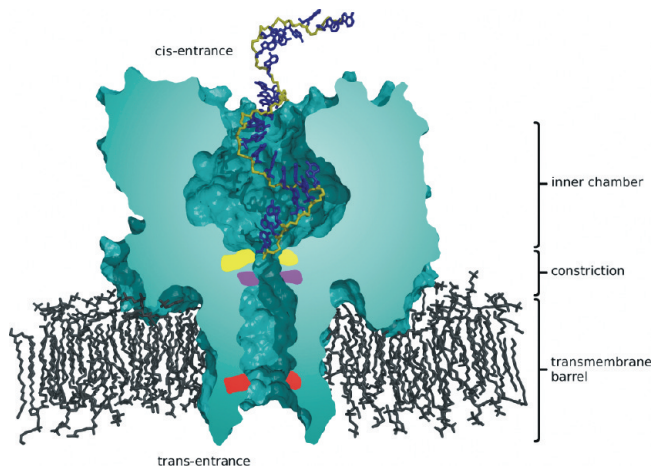


Figure 1. Figure representing the starting configuration of a 3' led A_{25} translocation simulation. The heptameric protein pore α HL (green) is inserted into a lipid bilayer (black). Features of the translocating molecule include the backbone of A_{25} (dark yellow) and the nucleic acid bases (blue). The *trans*-entrance is at the bottom of the pore; taking the *trans*-entrance of α HL as a reference point at 0 Å, other notable features include protein residue Leu-135 at 13 Å (red), Met-113 at 43 Å (pink), Lys-147 at 45 Å (light-yellow), and the *cis*-entrance at the top of the protein at 95 Å. The *cis*-entrance is 28 Å in diameter. The wide section of the pore running from the *cis*-entrance to residue Lys-147 is termed the inner chamber and is up to 46 Å wide. The constriction marked by residues Lys-147 and Met-113 is 14 Å wide, while the transmembrane barrel runs from the constriction to the *trans*-entrance and is around 20 Å wide. The *trans*-entrance is 24 Å wide. The C3' carbon atom of the 3' end nucleotide of A_{25} is aligned with the center of mass of the C_{α} atoms of protein residue 111, which lies at the mouth of the constriction, just above residue Lys-147. For the sake of clarity, water molecules and sodium and chloride ions are not displayed (they are found along the entire length of the pore).

pore that has been extensively studied in experiments and computer simulations.^{10–16,19–21} “Of particular note is the lumen and the two entrances of the protein pore. Going from the top to the bottom of the pore Figure 1, the lumen at the *cis*-entrance is 28 Å in diameter, widens at the inner chamber, and then narrows to 13 Å at the inner constriction. The diameter of the transmembrane barrel increased to 20 Å at the *trans*-entrance Figure 1. The narrow inner constriction and also the transmembrane barrel is the bottleneck for the transport of matter through the pore. We explore the protein pore translocation of the nucleic acid strands polyadenosine, poly(A), and polydeoxycytidine, poly(dC), which are single strands of RNA and DNA, respectively. Poly(A) and poly(dC) molecules of 100–200 bases in length exhibit a 20-fold difference in translocation time through α HL in SCCR experiments.⁷ In order to gain insight into the experimentally observed difference, we simulate translocation of these two polynucleotides using cv-SMD. We also translocate single nucleotides A_1 and dC_1 to discern the relative contributions to the free energy profiles. We then investigate aspects of the translocation process using site-specific mutations within α HL.

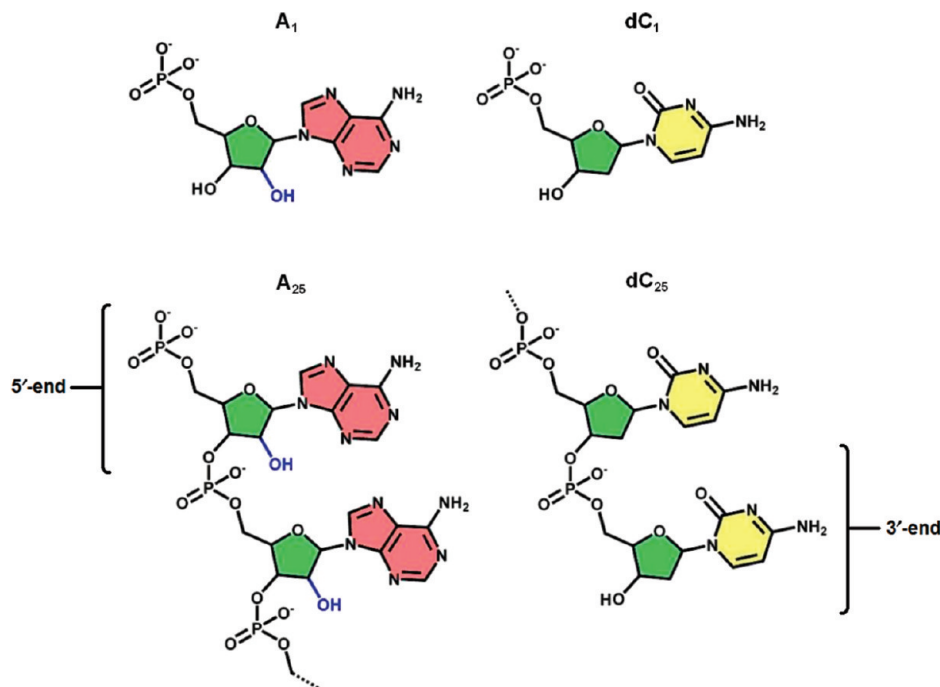


Figure 2. Molecular structures of A_{25} , A_1 , dC_{25} , and dC_1 . Structural features include the extra hydroxyl group associated with RNA (blue), the pentose base (green), the adenine base (light red) and the cytosine base (yellow). The 3' and 5' ends of the polynucleotides are also labeled for reference.

In the Methods section, we provide details of the model and techniques used to perform our simulations as well as some of the underlying theoretical formalism. In the Results and Discussion section, we present a detailed analysis of the results of our simulations; in the process, we discuss the optimization of various key parameters, visually inspect molecular configurations at equilibrium, and present the free energy profiles from polynucleotide and single nucleotide translocation through α HL in its wild and mutated forms. In the Conclusion section, we present our conclusions.

2. Methods

This section describes the computational construction of the transmembrane nanopore model and its associated components. We discuss Jarzynski's equality and explain our method for extracting free energy profiles from the simulations. We also describe the handling of errors for the simulations.

2.1. Model Construction. The α HL crystallographic structure coordinates were taken from Protein Data Bank (PDB) entry 7AHL. Atoms missing from the PDB file (residues *d*Lys-30, *g*Lys-30, *a*Lys-75, *d*Lys-240, *f*Lys-283, and *a*Arg-66) were reconstructed using psfgen, which is part of the NAMD2²² structure building module. α HL residues were mutated when required using the VMD²³ plug-in mutate. The protein was inserted into a patch of $150 \text{ \AA} \times 150 \text{ \AA}$ pre-equilibrated and solvated 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC) lipid bilayer using the VMD plug-in membrane, aligned to the *xy*-plane plane. The center of mass of the hydrophobic belt of α HL (residues 118–126 and 132–142) was aligned with the center of mass of the lipid bilayer. Lipid and water molecules overlapping the protein were removed. The system was solvated in a

water box of pre-equilibrated TIP3P²⁴ water molecules using the VMD plug-in solvate with a 2.4 \AA buffer distance between water molecules applied. The aqueous solution was set at 1 M NaCl using the VMD plug-in ionize. The dimensions of the lipid membrane and water box were chosen to minimize interactions between periodic images of the molecules due to the use of three-dimensional periodic boundary conditions. Figure 1 shows α HL inserted in a lipid membrane as in our models.

The protonation states chosen were consistent with SCCR recording pH range, typically between 7.5 to 8.5.^{6,11} The protonation states of the amino acid and phosphate groups were selected as follows: the amine groups of lysine and arginine residues were protonated, carrying a positive charge; interchain phosphate groups were unprotonated, carrying a single negative charge; and finally, terminal phosphate groups on the single nucleotide molecules were unprotonated, carrying a double negative charge.

The poly(A) and poly(dC) molecules (molecular structures depicted in Figure 2) were constructed using the AMBER module nucgen.²⁵ Nucleic acid polymers with 25 bases were constructed; the length was selected in order to maximize the influence of the nucleic acid polymer helix throughout the simulation, allowing comparisons to experimentally used lengths of 100–200 bases, while keeping the number of atoms in the model reasonably low. Double stranded conformations of the nucleic acid polymers were first built, the coordinates of which were derived from fiber diffraction studies.²⁶ A selected strand of the double helix was then removed from the PDB file, while the remaining single stranded nucleic acid polymer was charge neutralized with sodium counterions using the AMBER module xleap.²⁵ The nucleic acid polymer was then orientated as necessary using

the VMD module orient,²³ and the C3' carbon atom of the leading nucleotide was aligned with the center of the α carbon atoms (C_α) of protein residue 111. Water molecules that were overlapping the inserted polynucleotide were removed. After pulling the SMD atom from the constriction to the *trans*-entrance, the 25-base polynucleotide will have more than 10 bases remaining in the wide inner chamber. Our simulations with a solvated polynucleotide show that 10 bases are sufficient for it to form a helical structure in poly(A). An example of the starting position of the polynucleotides is shown in Figure 1.

Single nucleotide PDB files of adenosine monophosphate (A_1) and deoxycytidine monophosphate (dC_1) were also obtained (PDB identifiers AMP and DCM, respectively), and topology files were modified accordingly to produce accompanying PSF files. The molecular structures of these molecules are depicted in Figure 2.

The final models consisted of 328 000 and 262 000 atoms for the 25-base polynucleotide and single nucleotide models, respectively. The dimensions of the models were $145 \times 145 \times 154 \text{ \AA}$ for the 25-base nucleic acid polymer system and $145 \times 145 \times 121 \text{ \AA}$ for the single nucleotide system. The protein and nucleic acid molecules were at a minimum of 12 \AA from the edge of the simulation cell, ensuring a minimum distance of 24 \AA from the periodic images was maintained.

Translocation was chosen to be led by the 3' end (see Figure 2) of the polynucleotide for all the simulations in this paper. As shown by several groups, a difference is observed in the translocation times of nucleic acid polymers depending on whether they are threaded by the 3' or 5' end of the nucleotide chain.^{11,27,28} Nucleic acid bases projecting out along the phosphate backbone must flatten toward the backbone in order to pass through a tight constriction such as found in α HL. Nucleic acid bases tend to point toward the 5' end; therefore, they must bend through a larger angle than that of the 3' threaded polymers in order for the molecule to translocate through the pore. This results in an energetically less favorable process for 5' translocation than for 3' threaded instances.¹¹ Thus, 3' led translocation was selected to compare poly(A) and poly(dC). Given the nature of cv-SMD pulling, the stressing of bond angles, such as that noted for the deformation in the helical shape of the nucleic acid polymers, is likely to occur in the angle of the bases along the backbone. This influences the 5' led polymers more than the 3' ones as the bases could be made to invert orientation toward the 3' end by the force of translocation.

Translocation simulations in this paper were focused on SMD atom movement through the constriction and transmembrane barrel of α HL (see Figure 1). The dimensions of this part of the pore present the greatest resistance to a translocating molecule.¹⁹ The starting position of the 25-base polymer is shown in Figure 1; the SMD atom of the leading nucleotide is aligned with the center of the C_α atoms of the glutamic acid residue 111 at the top of the constriction. The dimensions of the inner chamber of α HL allow for the presence of a polynucleotide helix, whereas the smaller diameters of the constriction and transmembrane barrel do not allow it. Hence, at these constricting sections of the pore,

Table 1. The Translocation Molecules and Pore Types Simulated in This Work^a

system name	α HL type	nucleotide base	nucleotide units
A_1 -WT	Wild type	Adenine	1
A_{25} -WT	Wild type	Adenine	25
dC_1 -WT	Wild type	Deoxycytosine	1
dC_{25} -WT	Wild type	Deoxycytosine	25
dC_1 -Mut	Mutant	Deoxycytosine	1
dC_{25} -Mut	Mutant	Deoxycytosine	25

^a Wild type indicates α HL with no mutated residues, and mutant indicates α HL mutant L147M.

the helix of a translocating polynucleotide is required to unwind in order to pass through.

Constraints placed on the movement of protein amino acid residues are necessary in order to prevent the entire pore from being dragged downwards due to the cv-SMD applied force. To ensure that specific interactions between amino acids and nucleotides could take place, the side chains on the protein amino acids were left unconstrained to allow reorientation. Constraints of 0.5 N/m were, therefore, limited to the C_α atoms of the protein amino acid residues.

2.2. Simulation Details. Simulations were performed using the molecular dynamics simulation package NAMD version 2.6.²² NAMD is designed for scalable high-performance simulations on parallel machines. The CHARMM²⁹ force field was applied using all-hydrogen parameter files for CHARMM22 proteins and CHARMM27 lipids and nucleic acids. Periodic boundary conditions were applied, and the particle mesh Ewald method³⁰ was used to calculate electrostatic forces with a dielectric constant $\epsilon_0 = 1$ computed over $128 \times 128 \times 128 \text{ \AA}$ grids. Van der Waals energies were calculated with a smooth 10–12 \AA cutoff. Constant temperature was maintained at 295 K using Langevin forces.³¹ Constant pressure was maintained by Nosé–Hoover Langevin piston pressure control³² at 1.01325 bar. A 2 fs time step was used with SHAKE turned on. So as to eliminate high-energy atom clashes, models were energy minimized for 2 ps until the gradient tolerance was below 20. The temperature was raised by 5 K every 50 fs for 3 ps to a final temperature of 295 K. Prior to insertion of the nucleic acid, the α HL model was equilibrated in the NpT ensemble for 1.3 ns with the backbone of the protein constrained and for a further 4 ns without constraints. After insertion of the nucleic acid molecules, the model was equilibrated in the NpT ensemble for 4 ns with the alpha carbons of the protein constrained and the ribose unit's C3' atom of the leading nucleotide of the nucleic acid molecule fixed. Translocation was limited to 1 ns per simulation due to resource constraints and to control binning errors. An overlap of 0.2 ns between sequential simulations was performed to enable removal of startup artifacts.

In order to sample the ensemble, multiple simulations of the nucleic acid molecule translocating past the same section of the pore were required. The initial configurations used to perform these translocation samples were obtained by capturing snapshots of the atomic positions and velocities, separated by 0.2 ns at equilibrium, with the SMD atom position fixed and the C_α protein atoms restrained.

Table 2. Parallel Scaling Performance and Efficiency Data^a

processors	queen bee		HPCx	
	wall clock/ time step (s)	efficiency	wall clock/ time step (s)	efficiency
32	0.29	1.00	0.47	1.00
64	0.16	0.91	0.25	0.95
128	0.11	0.67	0.13	0.90
256	0.06	0.65	0.07	0.83
512	0.04	0.48	0.04	0.74

^a The performance is reported in terms of the wall clock time taken in seconds per timestep and the efficiency relative to 32 processor simulations for a 328 000 atom model. Efficiency, E , at i number of processors is calculated as $E(N_i; N_{32}) = (t_{32} N_{32}) / (t_i N)$, where N_i is the number of processors, N_{32} is the reference number of processors, t_{32} is the wall clock/timestep at 32 processors, and t_i is the wall clock/timestep using i number of processors. Data are presented for both the LONI/TeraGrid queen bee and HPCx machines.

Table 1 lists the key system configurations simulated. The molecular structures of the translocating molecules are depicted in Figure 2.

The simulations performed to generate the results we report here required more than 800 000 CPU hours. It is important to note that these were required on high-end tightly coupled machines, including those on HPCx (<http://www.hpcx.ac.uk/>), NGS (<http://www.grid-support.ac.uk/>), TeraGrid (<http://www.teragrid.org/>), and LONI (<http://www.loni.org/>). For the simulations reported here, we used between 256 and 512 processors per simulation. We list the parallel scaling performance and efficiency data for an atomistically detailed model containing 328 000 atoms in Table 2. See Supporting Information for an in-depth discussion of the computational resources used.

2.3. Jarzynski's Equality. In this work we have used Jarzynski's equality (JE)¹⁸ to extract free energies from the nonequilibrium simulations. JE equates the change in equilibrium free energy, ΔF , to the average ensemble work, W , by the following equality:

$$e^{-\beta\Delta F} = \langle e^{-\beta W} \rangle \quad (1)$$

Here, the angular brackets, $\langle \rangle$, indicate an ensemble average, and β is the inverse temperature, $1/k_B T$, where k_B is Boltzmann's constant. Thus, JE provides a means of equating the equilibrium free energy to the average work done in a nonequilibrium process.¹⁸

To derive Jarzynski's equality, consider a process changing a parameter, λ , of a system from equilibrium point 0 at a time zero to Λ at time τ . As a consequence of the second law of thermodynamics, the average work done, W , on the system cannot be smaller than the Helmholtz free energy (F) difference between the initial and final states of λ :

$$\Delta F = F(\lambda_i) - F(\lambda_0) \leq \langle W \rangle \quad (2)$$

Jarzynski showed that the free energy difference between the two states can be related to the work done to transition between them, W , by the equality shown in Equation 1. This is known as the Jarzynski's equality (JE).¹⁸ A rearrangement

of Equation 1 can be used to obtain the Helmholtz free energy difference in the form:

$$F_{\lambda(\tau)} - F_{\lambda(0)} = -\frac{1}{\beta} \log \langle \exp[-\beta W(\tau)] \rangle \quad (3)$$

The ensemble average of the exponential work ($\exp[-\beta W(\tau)]$) is taken over the ensemble of trajectories of the process. A problem in using Jarzynski's equality is that, due to the exponential, the free energy is dominated by small values of the work. In the absence of very extensive sampling, these small amounts of work are not included in the ensemble average of the exponential work. Therefore, instead of estimating the exponential of work directly, a cumulant expansion is often employed. The general form for the cumulant expansion of a logarithm of an ensemble average of an exponential (as in Equation 3) is

$$\log \langle e^x \rangle = \langle x \rangle + \frac{1}{2!} (\langle x^2 \rangle - \langle x \rangle^2) + \frac{1}{3!} (\langle x^3 \rangle - 3\langle x^2 \rangle \langle x \rangle + 2\langle x \rangle^3) + \dots \quad (4)$$

Inserting Equation 4 into Equation 3 gives

$$F_{\lambda(\tau)} - F_{\lambda(0)} = \langle W(\tau) \rangle + \frac{\beta}{2!} (\langle W(\tau)^2 \rangle - \langle W(\tau) \rangle^2) + \frac{\beta}{3!} (\langle W(\tau)^3 \rangle - 3\langle W(\tau)^2 \rangle \langle W(\tau) \rangle + 2\langle W(\tau) \rangle^3) + \dots \quad (5)$$

This series can be terminated at a chosen order to give an approximate formula for the Helmholtz free energy difference of the process. Using the cumulant expansion reduces the sampling error of estimating the ensemble average of the exponential directly. However, a truncation error is introduced by ignoring higher order terms. For this work, we initially established that a first-order cumulant expansion was valid (see Supporting Information for analysis comparing the first- and the second-order expansions).

The free energy as a function of a reaction coordinate is also known as the potential of mean force (PMF). For cv-SMD simulations, in order for the free energy profile to be expressed as a function of a well-defined reaction coordinate, a stiff spring is required to ensure that the SMD atom closely follows the constraint position.

2.4. Free Energy from cv-SMD Simulations. In this section we explain how the data output from cv-SMD simulations performed using NAMD are converted into free energy profiles using the JE. NAMD outputs the force (f) experienced by the harmonic spring connecting the SMD atom(s) and the constraint. The work can be calculated from the force according to the following force versus extension integral:

$$W = \sum_{i=1}^{N_S} f_i \Delta x_i \quad (6)$$

Here, Δx_i is the distance pulled at force f_i , and N_S is the number of intervals (bins) in the sum. In this paper, the work is calculated along the trajectory using a typical bin size $\Delta x = 0.25 \text{ \AA}$, enabling the resolution of well-defined peaks; as a noise reduction technique, we also use $\Delta x = 0.75 \text{ \AA}$.

The work can be plotted as an integral of the whole trajectory, which we refer to as a global plot, or each bin can be treated independently, representing the work done in that bin alone, which we refer to as a local plot. A global plot is, therefore, essentially a cumulative local plot.

Free energy profiles calculated from the average ensemble work are derived in principle by averaging over an infinite number of samples. In practice, they must be determined by a computationally feasible number of samples, while still effectively sampling the trajectory. The number of samples used in each free energy profile calculation is listed in the figures. As we show in sections 3.3, 3.4 and 3.5, the local representation of work and free energy shows peaks while the global representation gives a clearer comparison between the cumulative energy barriers to the translocation of two different systems, such as poly(A) and poly(dC). See Supporting Information for representations of the data obtained during the calculation process.

2.5. Error Handling. Just as in any experiment, multiple sources of error arise when using simulations to calculate free energies. In addition to the usual errors arising from numerical imprecision and approximations, there are errors due to the finite sampling used to compute ensemble averages and the choice of simulation parameters. Additional sources of error arise due to the finite approximations that are made in the expression used for computing the free energy as well as from the binning and smoothing of data. Errors can never be completely eliminated; for a fixed computational effort, at best, the errors can be controlled and their influence minimized. That is why, in this work, we have taken considerable care to discern the effect of the different types of errors, and we have planned our simulations accordingly so as to minimize their influence.

As described in the section 2.3, by increasing the force constant of the harmonic spring connecting the SMD atom and constraint position, the systematic errors due to deviations of the SMD atom from the reaction coordinate are reduced. However, a high spring constant also introduces statistical noise. An optimal balance between systematic and statistical error was found and is reported in section 3.1. Also, as discussed in section 2.3, the statistical error from applying JE, in its direct exponential form, is reduced by the use of the cumulant expansion formula, which introduces a systematic error due to truncation of the higher order terms. The systematic error can be controlled through this method to be smaller than the statistical error from exponential averaging, thus, providing a better overall estimate.

In determining the values of the work and energy averaged over multiple samples, there are two types of errors that must be accounted for. The first arises from the use of a coarse grained value for the work. The mean value is referred to as the bin value, and the error that arises from replacing the individual data points with the average taken over the bin, i.e., binning, is referred to as the bin error. The second error arises from taking the average of the bin value over multiple samples (sample-to-sample error); this error is statistical in nature and can be reduced with a greater number of samples. A challenge arises when the number of samples required to lower the statistical error is impractically large. Where both

error types arise, they are combined according the propagation of errors.

Additionally, as we explore in section 3.1, a high pulling speed introduces undesirable conformational changes. Thus, a balance must be found between performing a significant degree of translocation on a given resource and obtaining correct and accurate data.

3. Results and Discussion

In this section, we present the details of the optimization of parameters, free energy profiles, and other data from our MD simulations and the physical insight gained from them. In section 3.1 we describe the optimization of the cv-SMD spring constant and the chosen pulling speed. Section 3.2 presents the differences between the conformations of A₂₅ and dC₂₅, relating the variations to experimentally and computationally observed trends. Section 3.3 compares the free energy profiles for A₂₅ and dC₂₅ polynucleotide translocation. In section 3.4, we present the free energy profiles for A₁ and dC₁, comparing them to the polynucleotide free energy profiles. In section 3.5, we perform translocation through a mutated α HL pore, comparing the free energy profiles of the mutant and wild type pores.

3.1. Parameter Optimization. As described in section 2.1, there are many parameters that need to be carefully selected for effective cv-SMD simulations. For example, the choice of spring constant and pulling speed is especially sensitive to the size and conformational behavior of the pulled molecule, requiring explicit attention for the translocation of polymers. As demonstrated by Park and Khalili-Araghi,³³ a stiff spring connecting the SMD atom and the constraint position is required in order to use JE to extract the free energy correctly. The positions of the SMD atom and the constraint position may be significantly different with a low spring constant, causing greater deviation from the reaction coordinate. A stiff spring is essential; however, the statistical noise in the force output is higher for a larger force constant.

Figure 3a provides data which we use to find the optimized value of the spring constant for the translocation of dC₂₅ in wild type α HL. At a pulling speed of 0.04 Å/ps, spring constants 50, 100, and 200 kcal/mol are shown to approximate a stiff spring as the SMD atom closely follows the constraint position. A value of 100 kcal/mol was chosen to ensure the spring was stiff without introducing excessive statistical noise. Figure 3b shows that a 100 kcal/mol spring constant exhibits similar behavior at pulling speeds of 0.04 and 0.004 Å/ps. The sampling rate for the SMD force was every 40 fs at 0.04 Å/ps and every 400 fs at 0.004 Å/ps.

The statistical and systematic errors tend to increase as the SMD atom moves further from its starting position. Lower spring constants allow the SMD atom to lag behind the constraint position as the translocation length increases. As shown in Figure 3a, the position of the SMD atom for 1.4, 7, and 50 kcal/mol spring constants shows an increase in lag between the SMD atom and the constraint position at the end of 4 Å of translocation. Translocations with 100 and 200 kcal/mol spring constants show a negligible increase in lag at the end of 4 Å. The lag controlled by segmenting the

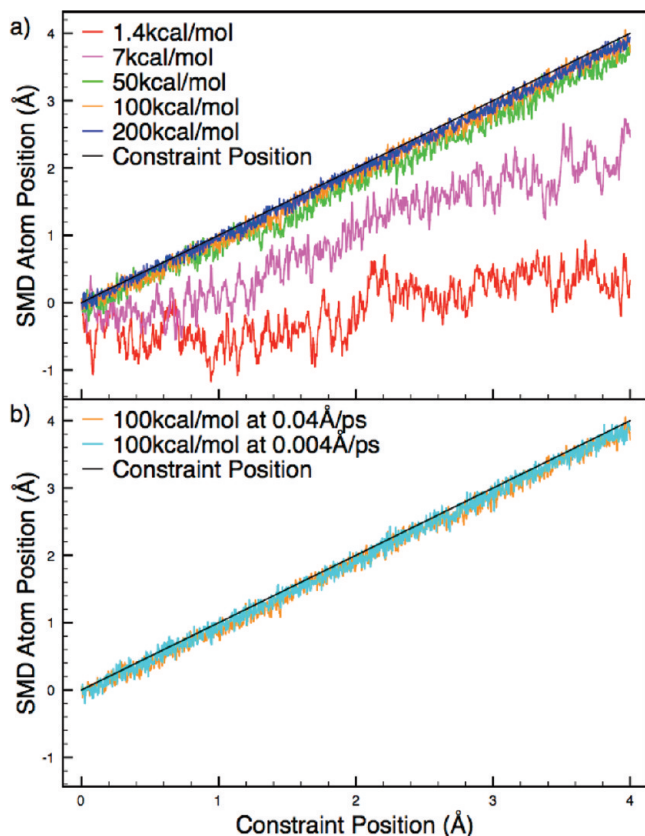


Figure 3. Analysis of the effect of spring constant values on the lag between the SMD atom and the constraint position. (a) The position of the SMD atom and the constraint position (Å) plotted against the constraint position for simulations with differing spring constants. The simulations were run at 0.04 Å/ps using spring constants of 1.4, 7, 50, 100, and 200 kcal/mol. The harmonic spring must be approximately stiff in order for the free energy to be accurately calculated along a well-defined reaction coordinate. This plot shows that the spring could be safely considered stiff at 100 kcal/mol, which was used for all subsequent simulations. (b) The SMD atom and constraint positions using a 100 kcal/mol spring constant pulled at 0.04 or 0.004 Å/ps show that the spring is approximately stiff at both speeds, with only a marginal increase in stiffness at the slower speed.

trajectory into 1 ns (4 Å) sections and piecing them together to form the full translocation length. Therefore, with a spring constant at 100 kcal/mol, the errors are controlled by resetting the constraint position to the SMD atom every 4 Å.

As shown by Wells et al.,¹⁶ if the system is too far from equilibrium, the translocation of a large deformable molecule in a cv-SMD simulation can result in undesirable conformational changes. This is due to translocation forces being applied to only a small part of a molecule. If the molecule moves too fast, then there is insufficient time for relaxation forces to return the structure to its equilibrium state. To reduce this effect, the speed of cv-SMD simulations for large molecules should be kept to a minimum. The challenge is to balance the desire for higher speeds with the need for correctness and accuracy. Figure 4 shows the change in separation between the first and second and the first and third nucleotides as the translocation of A₂₅ proceeds at various speeds. At slower speeds, the backbone has more time to

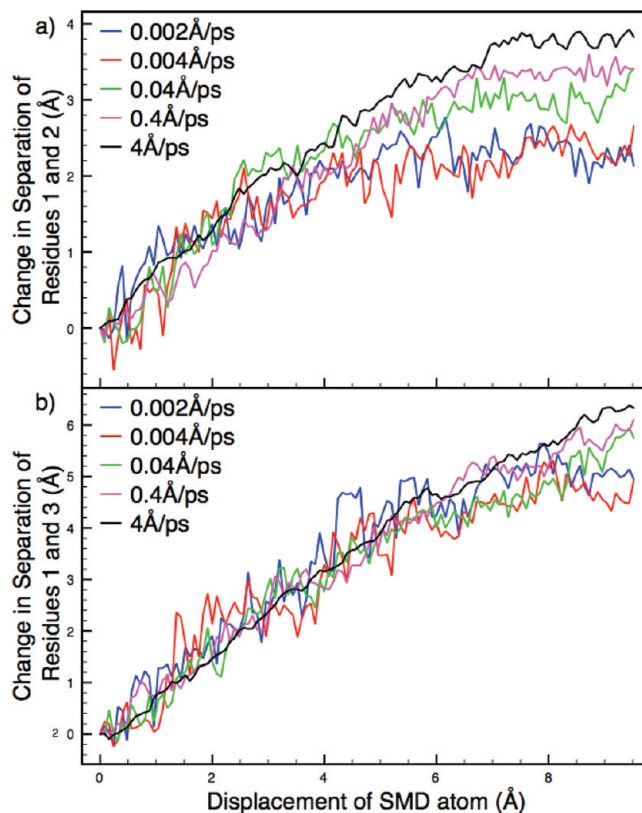


Figure 4. The effect of pulling speed on the relaxation of the conformation of translocating nucleic acid polymers. The plots show the change in z-axis separation between the pulled atom (C3' carbon atom of the front-most nucleotide) and the C3' atom of the (a) second and (b) third nucleotides during an A₂₅ pulling simulation at different speeds. At pulling speeds where the system is close to equilibrium (not shown), the separation between the first and other nucleotides in the polymer will not tend to increase. Both plots show that, at slower speeds, nucleotides further along the polymer from the pulling atom have time to relax conformation and to move closer to the pulled atom, reducing the stretching of the polymer. At higher pulling speeds, the polymer stretches more, making it desirable to keep the speed to a minimum. At a pulling speed of 0.002 Å/ps, the computational expense for our chosen translocation distances is too high; instead, 0.004 Å/s was chosen for simulations in this paper.

relax from deviations from its normal conformations as a result of the pulling, thereby keeping nucleotides closer to each other during translocation. The increased separation is more pronounced as the nucleotide under consideration is further along the chain from the pulled atom, highlighting the need to keep the pulling speed to a minimum. Taking these factors into account, a speed of 0.004 Å/ps was selected for the simulations reported in this paper, which is roughly a 1 000 times faster than translocation speeds found in experiment. While Akesson et al.⁷ hypothesized that the unwinding of a poly(A) helix gives rise to a long translocation time, it should be noted that at a translocation speed of 0.004 Å/ps, it is difficult to preserve the helical structure of the polynucleotide in the nonconstricting inner chamber.

3.2. Base Stacking and Secondary Structure. In this section, we examine the conformations of A₂₅ and dC₂₅ at equilibrium. Our simulations are consistent with the predic-

tion that poly(A) exhibits a greater degree of base stacking than poly(dC) molecules,^{34–37} which is thought to contribute to the experimentally observed 20-fold higher translocation time of poly(A) compared to poly(dC).⁷ Poly(A) possesses a helical structure with a diameter that is greater than the constriction and transmembrane barrel of α HL. According to Akesson et al.,⁷ the helix is believed to unwind during translocation, posing an energy barrier. It is known that overlap of the aromatic π -orbitals in the nucleic acid bases causes them to stack, giving rise to the helical conformation of the backbone. While the CHARMM27 forcefield does not explicitly describe π -orbital overlap, it accounts for the base stacking that results from the interactions between aromatic groups. “[Base-stacking occurs due to the overlap of the aromatic π -orbitals, therefore the two aromatic rings on purine bases (such as adenine) give rise to a greater degree of stacking compared to that of the one aromatic ring found in pyrimidine bases (such as cytosine).] While the CHARMM27 forcefield does not explicitly describe π -orbital overlap, it accounts for the base stacking that results from the interactions between aromatic groups. Polynucleotides can adopt a helical shape as a result of base-stacking, thus poly(A) is expected to be helical in solution, whereas poly(dC) is expected to take a more random conformation. The predicted poly(A) helical structure has a diameter that is greater than the constriction and transmembrane barrel of α HL. According to Akesson et al.,⁷ the helix is believed to unwind during translocation, posing an energy barrier, while poly(dC) experiences a lower activation barrier to unwinding and translocates through constricting pores faster.”

At equilibrium, our systems containing poly(A) or poly(dC) exhibit a difference in mean conformation. As can be seen from Figure 5, which shows the average structure during a 2.6 ns simulation at equilibrium, A_{25} displays a helical shape and the bases align in a well stacked formation. On the other hand, dC_{25} displays a less defined helical backbone with reduced stacking of its bases.

As reported by Wells et al.,¹⁶ previous MD simulations indicate that base stacking can influence the translocation time of nucleic acid polymers. These authors found that a break in the base stacking passing through the constriction caused a temporary slowing down in translocation due to steric obstruction. Our own explorative simulations of stretched and unstretched polynucleotides indicate that base stacking can occur even in polynucleotides with a linear conformation. Therefore, energy barriers associated with the rupture of base stacking do not necessarily imply that helical unwinding has taken place. Base stacking and helical unwinding may, thus, represent two energetic barriers to translocation.

While the average structures show a difference in base stacking and conformations of A_{25} and dC_{25} , we will find in sections 3.3 and 3.4 that peaks in the local work and free energy profiles do not appear to correlate to stacked or unstacked sections along A_{25} or dC_{25} .

3.3. Simulations of Polynucleotide Translocation. We present the free energy profiles from the translocation of 25 base poly(A) and poly(dC) nucleotides (systems A_{25} –WT

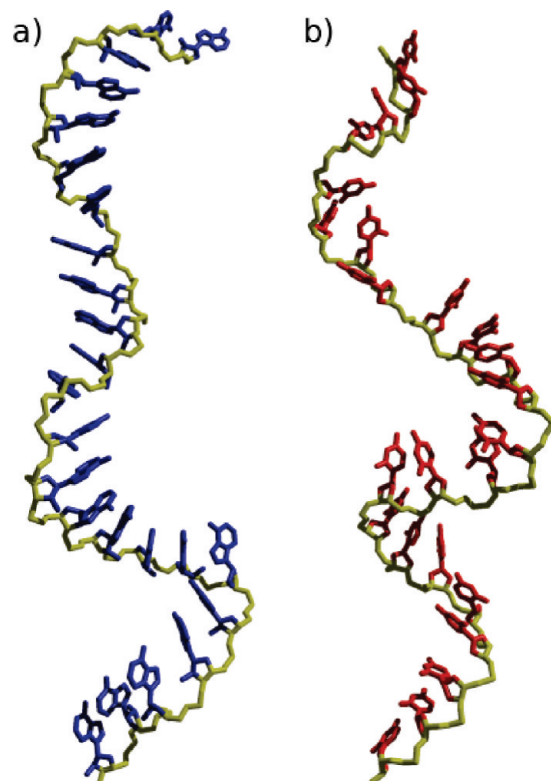


Figure 5. The average structures of (a) A_{25} and (b) dC_{25} over a 2.6 ns period at equilibrium in systems A_{25} –WT and dC_{25} –WT, respectively (see Table 1). The C3' atom of the front-most nucleotide is fixed in the center of the C_{α} atoms of the protein residue 111 at the top of the α HL constriction. The polynucleotides lie in both the inner chamber and outside the *cis*-entrance of α HL as shown in Figure 1. The A_{25} bases stack to a high degree, resulting in the more pronounced helical shape of its backbone than in dC_{25} . The dC_{25} structure does exhibit base stacking, though this is significantly less than that of A_{25} and is reflected in the more random configuration of the backbone. The increased stacking in A_{25} is due to the greater number of aromatic π -orbitals in adenine compared to cytosine, resulting in a more favorable π -orbital overlap.

and dC_{25} –WT (see Table 1)). The two polymers were shown, in the previous subsection 3.2, to exhibit different conformational behavior at equilibrium. We investigate the use of cv-SMD to produce free energy profiles from the translocation of these polymers in order to determine whether the profiles reflect experimentally determined longer translocation time for poly(A) compared to poly(dC).⁶ In addition, the simulations were conducted to gain physical insight into the differences.

The free energy profiles for a translocation length of 48 Å are shown in Figure 6 for A_{25} and dC_{25} with two samples performed for each trajectory system. Here, the SMD atom at the 3' end of the nucleic acid polymer was pulled from the top of the constriction to the bottom of the transmembrane barrel. The pore dimensions, as listed in Figure 1, indicate that the steric barriers to translocation occur mainly within this region. See Supporting Information for snapshots of the polynucleotide conformations during such a cv-SMD simulation.

Figure 6a shows the local free energy profiles of A_{25} and dC_{25} . Taking error bars into consideration, the two profiles

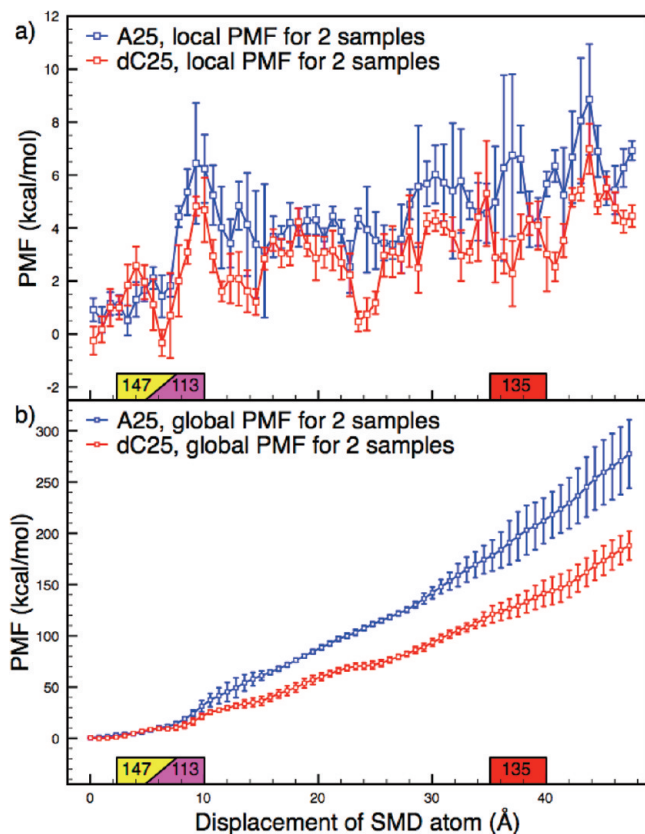


Figure 6. (a) Local free energy profiles of A_{25} and dC_{25} translocation from the top of the constriction to the bottom of the *trans*-entrance; each profile was derived from two samples. Labeled along the *x*-axis are protein residues Met-147, Lys-113, and Leu-135. The residue labels span 5 Å from the pulled atom to when first phosphate atom passes the labeled residue. The plot shows poor separation of A_{25} and dC_{25} when considering the error bars, though a general trend, where A_{25} has a higher free energy profile than dC_{25} is observed. An overall increase in the free energy profile is observed from left to right, which is expected as more nucleotides enter the confining dimensions of the constriction and transmembrane barrel. (b) Global free energy profile of A_{25} and dC_{25} translocation from the top of the constriction to the bottom of the *trans*-entrance; each profile was derived from two samples. The plot shows good discrimination of A_{25} and dC_{25} beyond the error bars.

are poorly separated when plotted in the local form, although A_{25} tends to exhibit a higher free energy than dC_{25} . Large peaks found in the local work profiles of individual samples are not as pronounced in the local free energy profiles consisting of multiple samples. The size of these peaks varies considerably from sample-to-sample and, other than the peak at 9 Å, their positions also vary.

The global free energy plots for the 48 Å translocation of A_{25} and dC_{25} are shown in Figure 6b. A_{25} displays a higher free energy profile than dC_{25} , with the discrimination well beyond the error bars. The separation between the means of the profiles continues to grow throughout the translocation process. The error bars are larger for A_{25} than for dC_{25} suggesting a higher sample-to-sample variation in the former case as the binning errors represent a small portion of the total errors. The higher free energy profile for A_{25} compared

to dC_{25} is in qualitative agreement with the longer experimental translocation times for A_{25} .⁷

Jensen et al. estimated the mean first passage time of glycerol translocation through the protein pore aquaglyceroporin GlpF from the reconstructed free energy difference of the translocation event.³⁸ This method relates the translocation time to the integral of the exponential free energy difference along the reaction coordinate, with constants representing the friction coefficient and inverse temperature. Using this method, we compute the relative translocation times of A_{25} and dC_{25} from the data in Figure 6b over the complete range. By assuming that the friction coefficients of A_{25} and dC_{25} translocation through the pore are similar, we find that the estimated translocation times have an approximate ratio of 25:1. Due to the exponential dependence of the translocation time on the free energy difference, reliable estimates of the translocation time without precise parameters such as friction coefficients are difficult. However, it is worth noting that our analysis is consistent with the 20:1 translocation time ratio of poly(A) and poly(dC) found in experiments.⁷

Figure 7 shows the free energy profiles of smaller segments of the trajectory, constructed from more samples to represent the ensemble average. Two parts of the pore are examined in this way, chosen to exhibit high energy barriers in the 48 Å translocation: the constriction, which is around residues Met-113 and Lys-147 (Figure 7a and b), and the transmembrane barrel, specifically around residue Leu-135 (Figure 7c and d). The diameter of the constriction is smallest at residues Lys-147 and Met-113, while the transmembrane barrel is smallest at leucine residue Leu-135; these residues are marked in Figure 1.

The free energy profiles in Figure 7 tend to show an increased energetic barrier to translocation of A_{25} relative to dC_{25} . Due to high variation in the single sample work profiles, the local free energy plots exhibit large overlapping error bars between A_{25} and dC_{25} (Figure 7a and c). The high variation in the single samples also leads to less pronounced peak shapes in the local free energy profiles. Exploration of peaks and local barriers to translocation is, therefore, best performed through examination of the local work profiles from single samples (see further below). In contrast to the local profiles, the global profiles show good separation between A_{25} and dC_{25} around protein residue Leu-135 (Figure 7d) and poor separation at the constriction (Figure 7b), although they point toward A_{25} exhibiting a higher free energy profile by the end of the trajectory. The error bars in the global free energy profiles again show greater sample-to-sample variation for A_{25} than for dC_{25} (Figure 7c and d).

We explored the molecular reason for the high peaks in the profiles of single samples. We visually inspected a molecular model of A_{25} at the inner pore constriction and observed that the phosphate groups of the polynucleotide strand interacts electrostatically with the positively charged amine group of residue Lys-147 (Figure 8). See Supporting Information for further investigation of the phosphate–lysine interaction.

The local single sample work profiles (not shown) contain higher peaks correlating to electrostatic phosphate–lysine

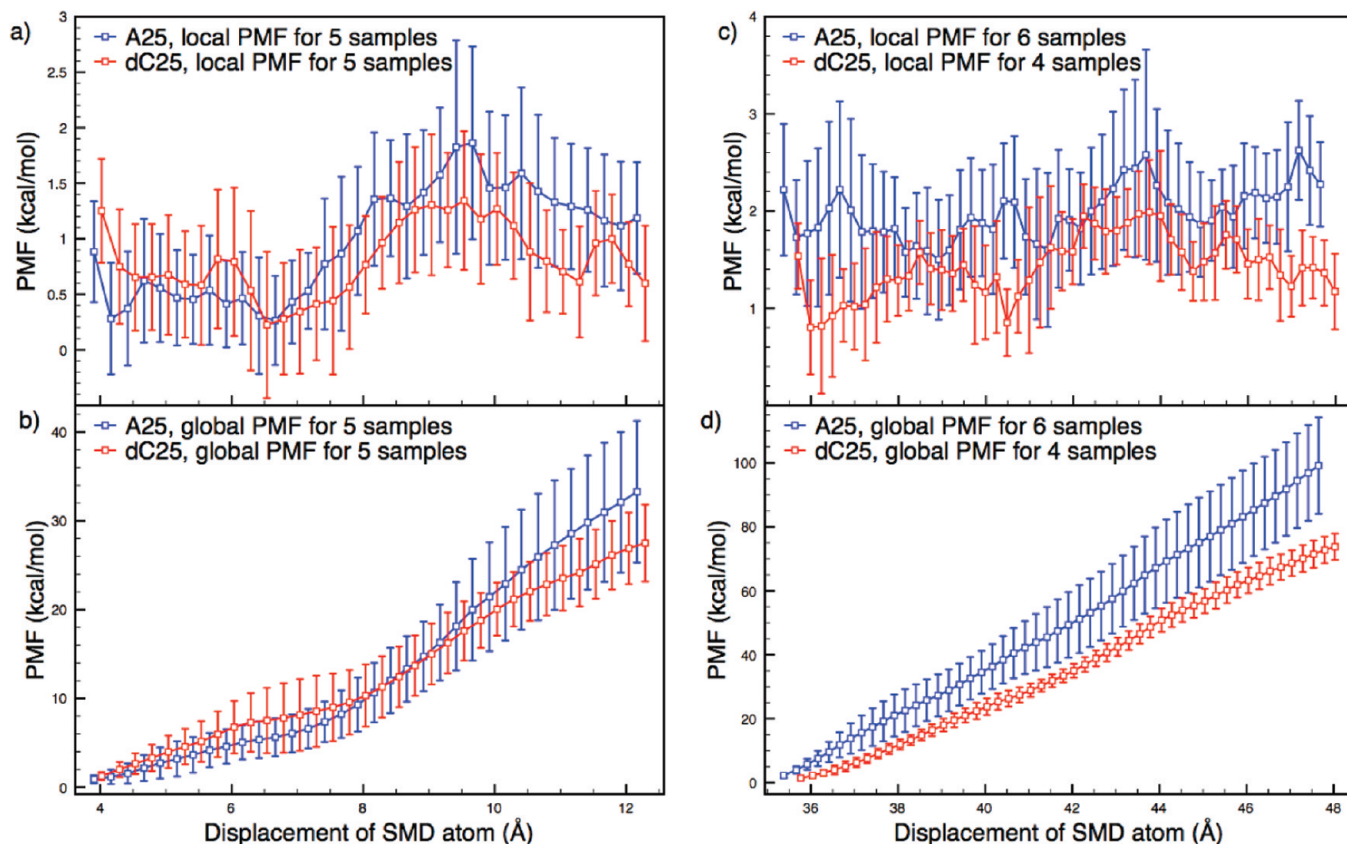


Figure 7. Local and global free energy profiles at the constriction and in the transmembrane barrel with increased sampling. (a) Local free energy profiles of A₂₅ and dC₂₅ translocating at the constriction. The statistics are based on 5 samples for each A₂₅ and dC₂₅. The separation is not beyond the error bars though both show a broad peak between 7 and 12 Å. (b) Global free energy profiles of A₂₅ and dC₂₅ translocating at the constriction. These show a tendency toward discrimination between A₂₅ and dC₂₅, albeit not beyond the error bars. (c) Local free energy profiles of A₂₅ and dC₂₅ translocating in the transmembrane barrel; there are 4 samples for dC₂₅ and 6 samples for A₂₅. The free energy of A₂₅ and dC₂₅ are the same to within errors. (d) Global free energy profiles of A₂₅ and dC₂₅ translocating in the transmembrane barrel. These show separation of A₂₅ and dC₂₅ beyond the error bars. The error bars in both global-plots indicate the greater variation between samples of A₂₅ compared to those of dC₂₅, indicating a higher degree of sample-to-sample variation.

interactions for A₂₅ than for dC₂₅. This suggests a more accessible and/or a more electrostatically charged phosphate group on A₂₅. Base stacking or base size may play a role in making the A₂₅ more accessible by pushing the phosphate group closer to the sides of the pore. Figure 2 shows the molecular structures of the translocating molecules, 2' hydroxyl group shown on the pentose sugar of A₂₅ may contribute to the electrostatics and sterics of the phosphate group. This hydroxyl group is inherent to RNA polymers such as A₂₅, and exhibits an electrostatic/steric repulsion between the hydroxyl group and the 3' attached phosphate group. See Supporting Information for further investigation of the effect of the RNA–hydroxyl group on the polynucleotide phosphates.

3.4. Simulations of Single Nucleotide Translocation.

In this section we present the free energy profiles from single nucleotide translocation in systems A₁–WT and dC₁–WT (see Table 1). The free energy profiles derived from polynucleotide translocation presented in the previous subsection represent combined energy barriers from the interactions of multiple nucleotides with the protein pore. To gain a clearer picture of the contributions from specific interactions, we performed translocation simulations on single nucleotides, and we present the resulting free energy profiles

here. By obtaining both the polymer and single nucleotide profiles, physical insight can be inferred with greater confidence.

As shown in Figure 9 for both A₁ and dC₁, the majority of activity occurs in the constriction region between 2 and 12 Å. For A₁, the increase in the free energy profile is broad, whereas for dC₁, there is a large, sharp increase between 8 and 12 Å. Inspection of the molecular models around the sharp peak in dC₁ shows that the increase clearly correlates with a strong electrostatic interaction occurring between the phosphate group of dC₁ and the amine group of Lys-147 (see Supporting Information for an animation of this interaction).

Comparing the molecular models of A₁ and dC₁ during equilibration and translocation, we find a clear difference in the local sodium ion positions. During equilibration of A₁ and dC₁ (not shown), sodium ions can be seen to approach both molecules. A sodium ion approaches the A₁ molecule, associates itself between the phosphate and the purine base, and remains there for the rest of the equilibration as well as for the start of the translocation simulations. For dC₁, sodium ions approach the molecule but do not appear to associate with the pyrimidine base, moving away before the end of the simulation. During translocation (not shown), the A₁

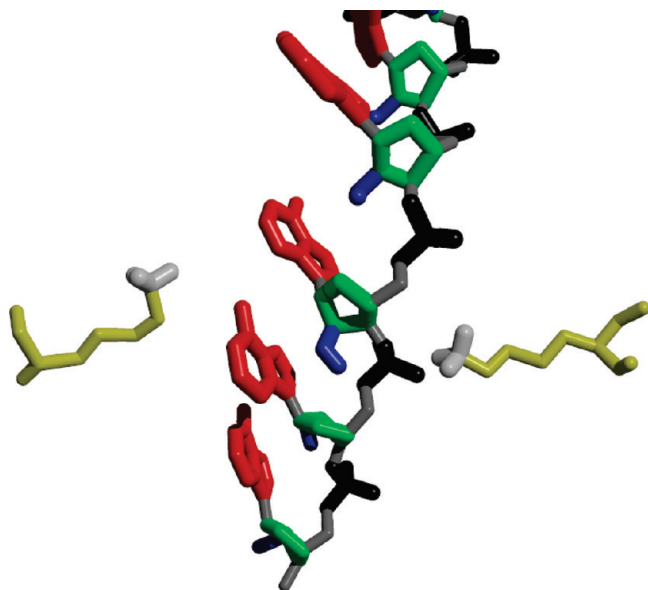


Figure 8. Molecular representation of the interaction between the nucleic acid phosphate groups of A₂₅ and the amine group of residue Lys-147. The polynucleotide is shown in the middle with a lysine amino acid shown from either side of the constriction. The hydrogen atoms, other than those on the amine of lysine and the 2' hydroxyl groups, have been omitted for clarity. Notable features include the polynucleotide's adenine group (red), pentose base (green), and the hydroxyl group (blue), which are colored to match Figure 2. Also shown is the lysine residue (yellow), its amine group (white), and the polynucleotide's backbone phosphate groups (black). The lysine group nearest the phosphates of the polynucleotide shows a reorientation of its side chain toward the phosphate, while the lysine on the other side remains in a conformation that is observed during equilibration.

associated sodium ion stays close to the phosphate group until the phosphate group moves past residue Lys-147; at this point the sodium ion remains above the constriction, while A₁ translocates down the pore. By contrast, dC₁ does not possess an associated sodium ion at any point during the translocation. This sequence of events occurred in all instances of A₁ and dC₁ translocation that was performed, though further sampling of these models would be required in order to draw strong conclusions about the influence of the aromatic bases on the associated ions. Aromatic groups have a known tendency to form strong interactions with cations,³⁹ just as we observe for A₁. Intuitively, one would expect the interacting sodium ion of A₁ to diminish the effective charge on the phosphate group and, thereby, reduce the interaction with the positively charged amine group of lysine. By contrast, it is anticipated that the uncoordinated phosphate of dC₁ will interact strongly with the lysine side chain, as is apparent in the sharp and repeatable peak in the energy profile in Figure 9a. See Supporting Information for an animation of the A₁ associated ion during translocation.

Figure 10 represents the global free energy profiles for the translocation of A₁ and dC₁. The figure shows that the sample-to-sample variance is much greater for A₁ translocation than for dC₁. The error bars in Figure 10 are roughly 4–5 times larger for A₁.

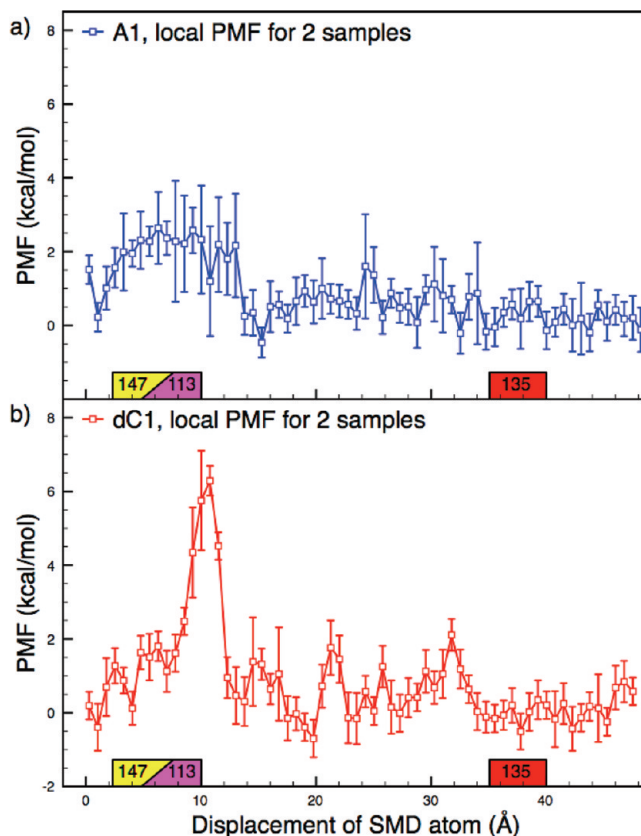


Figure 9. Local free energy profiles for the A₁ and dC₁ translocation, from the top of the constriction to the bottom of the transmembrane barrel, are derived from two samples. (a) Local free energy profile for A₁. A broad peak is shown between 2 and 12 Å which corresponds to the constriction of αHL. The error bars in this region are large, indicating a high degree of sample-to-sample variation. There is little notable activity exhibited in the transmembrane barrel. (b) Local free energy profile for dC₁. A sharp peak is observed at 10 Å, and an examination of the model at this point reveals a strong interaction occurring between the phosphate group of dC₁ and the amine group of residue Lys-147 at the top of the constriction. The error bars in this region are small, indicating a low degree of sample-to-sample variation. Examination of this region in A₁ shows some degree of interaction between the phosphate group of A₁ and the amine group of residue Lys-147, though it is considerably reduced compared to dC₁.

By comparing the gradients of the global energy profiles from the single nucleotide and nucleic acid polymer simulations (Figures 10 and 6b), contributions to the free energy profile from the constriction and transmembrane barrel can be deduced. Of note is the high gradient beyond 30 Å in the polynucleotide global plots compared to considerably lower gradients at this stage in the single nucleotide simulations. The low profile gradients for single nucleotide translocation in the transmembrane barrel give an indication of the energy barriers experienced by the leading nucleic acid nucleotide for polynucleotide translocation in this region. Therefore, the high gradients for transmembrane polynucleotide translocation indicate that increased resistance to translocation is occurring further along the polymer as other nucleotides are passing through the constriction.

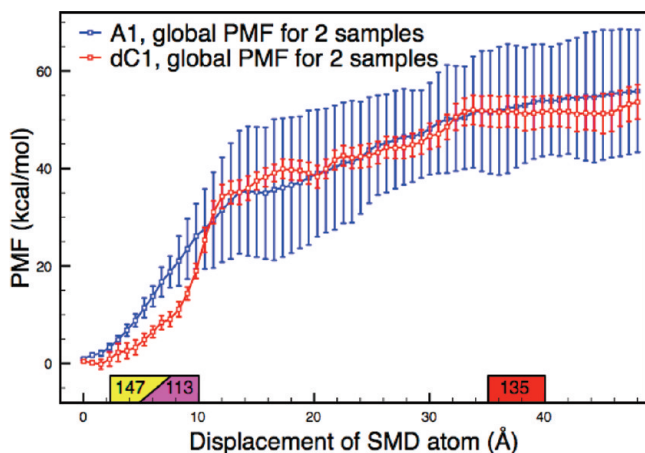


Figure 10. Global free energy profiles of the A_1 and dC_1 translocation, from the top of the constriction to the bottom of the transmembrane barrel, are derived from two samples. The gradients of the plots show that the majority of activity for single nucleotides occurs at the constriction. The plots are similar throughout the translocation process, showing separation outside of the error bars at the start of the constriction. For the majority of the trajectory, the error bars are larger for A_1 than for dC_{25} by 4- to 5-fold, showing a larger sample-to-sample variance for A_1 , which is in line with findings for the polynucleotide simulations.

3.5. Effect of α -HL Mutation. In this section, we present the free energy profiles from polymer and single nucleotide translocation in the dC_1 -Mut and dC_{25} -Mut systems (see Table 1). From the single nucleotide and polynucleotide simulations in the unmutated α HL pore (see Table 1, wild type systems), the phosphate-lysine interaction appears to play a significant role. To confirm that this interaction gives rise to the observed peaks in the free energy profiles, we perform nucleotide translocation through an α HL pore where the residue giving rise to the interaction is mutated. Figure 11 compares the free energy profiles from translocation through the mutant to that of the wild type pore.

Figure 11 shows the local and global free energy profiles from dC_1 and dC_{25} translocation through wild type α HL and α HL where residue Lys-147 is mutated into methionine. Methionine is similar in size to lysine. Therefore, the mutation is not expected to alter the energetic barriers of the constriction beyond removal of the charged amine group on lysine. The free energy profiles show a profound difference between the wild type and mutant for dC_1 translocation in both the local and global forms of the profile (Figure 11a and b), confirming that the phosphate-lysine interaction, indeed, gives rise to the peaks in the wild type profiles. dC_{25} shows a similar global profile for the wild type and mutant (Figure 11c and d); while a major difference is not observed in the local profile, a change in shape comparable to that seen in dC_1 is observed. As shown in previous sections, dC_{25} exhibits the weakest phosphate-lysine interaction of the four nucleotides investigated in this paper, resulting in smaller peak sizes in the free energy profiles. This accounts for the reduced differentiation between the global plots of dC_{25} translocation in the mutant and wild-type pores. Initial attempts at A_1 and A_{25} translocation through α HL mutant L147 M did not show a significant

difference in the free energy profiles compared to that of the translocation through wild type α HL.

The degree to which the phosphate-lysine interaction plays a role in single channel current recording experiments is unknown. Factors indicating that the interaction could be significant include the observation that there are approximately 12 positively charged protein residues accessible to a molecule translocating in α HL and that current-driven translocation is about 1000 times slower than in the simulations reported here. Conversely, under conditions pertaining in experimental translocation experiments there exists a rapid ionic current flow, not present in our simulations, which may reduce the influence of the electrostatic interaction. In SCCR experiments, it has been found that increasing the number of positive charges inside the α HL pore increases the capturing frequency of polydeoxyribonucleotide translocations,²⁰ indicating an electrostatic relationship between positively charged residues and negatively charged translocating polynucleotides. Our simulations show a high sensitivity to local ion positions, giving rise to a greater sample-to-sample variance in the free energy profiles. Therefore, in order to draw comparisons between noncurrent driven translocation simulations and SCCR experiments, careful consideration of the role played by ions is required.

4. Conclusions

Jarzynski's equality (JE) in conjunction with constant velocity-steered molecular dynamics (cv-SMD) has been used to compute the free energy profiles from atomistic molecular dynamics simulations of nucleic acid translocation through the protein nanopore α -hemolysin. Obtaining the free energy profiles for such large atomistically detailed systems, including long deformable translocating molecules, is unprecedented.

We present the free energy profiles from the translocation of adenine and deoxycytosine based polynucleotides. The results are in qualitative agreement with experimental data, showing poly(A) experiences a greater resistance to translocation through α HL than poly(dC). By obtaining the translocation free energy profiles for the translocation of single nucleotides, we show that the majority of peaks representing significant barriers to translocation occur as a result of interactions with the inner constriction. The single nucleotide profiles exhibit stronger electrostatic interactions for deoxycytosine than for adenine, contrary to the polymer translocation profiles. Through inspection of the molecular positions and conformations and by plotting the radial distances of phosphate and lysine groups, we are able to show that a phosphate-lysine interaction between the protein and nucleic acid makes a dominant contribution to the translocation energy barrier. By mutating the interacting lysine group into a methionine (L147M), we are able to confirm that the phosphate-lysine interaction is, indeed, responsible for sharp peaks in the free energy profiles.

By examination of the molecular structures and investigation of the contributions of local solvated cations we can account for many of the patterns observed in the free energy profiles. We attribute the stronger phosphate-lysine interactions in systems containing poly(A) to a more electrostatically charged and sterically accessible phosphate group on the polyribonucleotide. This occurs due to an

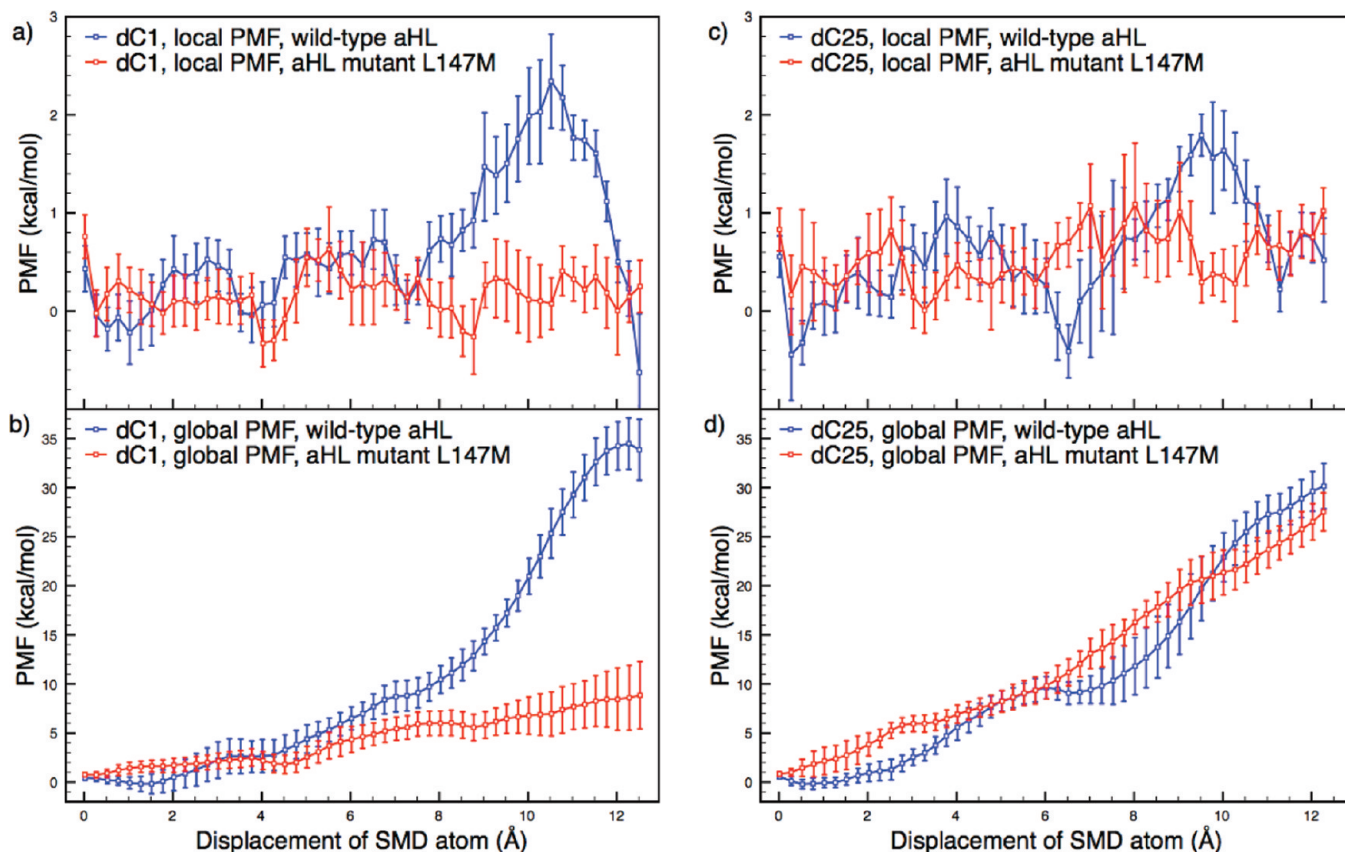


Figure 11. Free energy profiles for dC_1 and dC_{25} translocation, through wild type α HL and mutant L147M, are derived from two samples. Residue Lys-147 was mutated due to an observed interaction with dC_1 that appeared to cause a strong peak in the local free energy profile. (a) Local free energy profiles for dC_1 translocation through the constriction of wild type α HL and mutant L147M. The profiles show that, with the lysine residue mutated, the interaction does not occur, and the peak is not present in the free energy profile. (b) Global free energy profiles for dC_1 translocation through the constriction of wild type α HL and mutant L147M; the difference between the two models is represented well in this global form. (c) Local free energy profiles for dC_{25} translocation through the constriction of wild type α HL and mutant L147M. The local profiles do not manifest a good separation between the wild type and mutated pore but do show different shapes. (d) Global free energy profiles for dC_{25} translocation through the constriction of wild type α HL and mutant L147M. Here the global profiles do not show a discernible difference between the two pores. The peak present between 8 and 12 Å in for dC_1 -WT is also seen for dC_{25} -WT at the same position, though somewhat less pronounced. Simulations of the translocation of A_1 and A_{25} through the α HL mutant L147 M did not show a significant difference in the free energy profiles compared to translocation through the wild type pore.

energetically unfavorable interaction between the phosphate group at the 3' carbon and the neighboring 2' hydroxyl group. The hydroxyl group is not present in polydeoxyribonucleotides such as poly(dC). The higher sample-to-sample variation we observe in the free energy profiles of poly(A) translocation is due to variation in the local cation position combined with the enhanced sensitivity of poly(A) relative to poly(dC) due to its greater electrostatic charge.

This relationship between the phosphate and hydroxyl group does not occur in the nucleotide monomers due to the much larger distance between the phosphate at carbon 5' and the hydroxyl group at carbon 2'. For the translocation of adenine and deoxycytosine monophosphates, we find that the combination of the negatively charged phosphate group and the more aromatic adenine base causes a strong association with sodium ions, leading to a reduced phosphate-lysine interaction in adenine monophosphate. For deoxycytosine monophosphate translocation, the less aromatic cytosine base does not associate with local sodium ions, leading to an unbound phosphate group

and a stronger phosphate-lysine interaction. By providing evidence of the importance of phosphate-lysine interactions, our simulations set the stage for experimental confirmation.

In the future, it would be interesting to take advantage of the advent of even more substantial – petascale – computational resources⁴⁰ to examine slower translocation speeds and higher sampling for the free energy profiles. A comparison of poly(A) to poly(dA) would help to clarify the effect of the RNA-hydroxyl group on the phosphate-lysine interaction. Such data would illuminate the contribution of steric friction on translocation free energy profiles under these conditions. Single channel current recording data on an L147 M α HL mutant would indicate the influence of the phosphate-lysine interaction in experiments.

Overall, our study has yielded free energy profiles for translocation using constant velocity-steered molecular dynamics at speeds considerably higher than experiment, but nonetheless provides detailed explanations for the differences observed between selected nucleotide molecules. The findings furnish novel explanations for experimentally observed

translocation time differences. While our simulations are not performed under ideal conditions for close comparison with experimental results, they represent the state-of-the-art achievable with existing high-end computing resources and point the way to future enhancements that will become possible at the petascale and beyond.

Acknowledgment. We thank Daniel Keedy for his support with force field parametrization of the single nucleotide models. This work was supported by the U.K. Engineering and Physical Science Research Council (EPSRC) through the RealityGrid project (GR/R67699, EP/C536452/1), which also provided access to HPCx (<http://www.hpcx.ac.uk/>). TeraGrid (<http://www.teragrid.org/>) access was made possible through the National Science Foundation MRAC grants TG-DMR070014N and TG-DMR070013N and DAC grant TG-ASC070019T; significant computational resources on LONI were also made available via a grant to S.J. We are grateful to EPSRC and UCL Department of Chemistry for funding H.S.C.M.'s Ph.D. studentship.

Supporting Information Available: Discussions of computational resources utilized, Jarzynski's equality, snapshots of polynucleotide conformations during translocation, further investigation of the phosphate–lysine interaction, investigation of Crooks fluctuation theorem alongside JE, animations related to single nucleotide translocation, and a listing of significant residues in the α HL pore. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- Mullen, C.; Kilstrup, M.; Blaese, R. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 33–37.
- Davies, J. *Science* **1994**, *264*, 375–382.
- Koonin, E.; Makarova, K.; Aravind, L. *Annu. Rev. Microbiol.* **2001**, *55*, 709–742.
- Dreiseikelmann, B. *Microbiol. Mol. Biol. Rev.* **1994**, *58*, 293–316.
- Hanss, B.; Leal-Pinto, E.; Bruggeman, L.; Copeland, T.; Klotman, P. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 1921–1926.
- Kasianowicz, J.; Brandin, E.; Branton, D.; Deamer, D. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 13770–13773.
- Akeson, M.; Branton, D.; Kasianowicz, J.; Brandin, E.; Deamer, D. *Biophys. J.* **1999**, *77*, 3227–3233.
- Meller, A.; Nivon, L.; Brandin, E.; Golovchenko, J.; Branton, D. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 1079–1084.
- Meller, A.; Nivon, L.; Branton, D. *Phys. Rev. Lett.* **2001**, *86*, 3435–3438.
- Butler, T.; Gundlach, J.; Troll, M. *Biophys. J.* **2006**, *90*, 190–199.
- Mathe, J.; Aksimentiev, A.; Nelson, D.; Schulten, K.; Meller, A. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 12377–12382.
- Clarke, J.; Wu, H. C.; Jayasinghe, L.; Patel, A.; Reid, S.; Bayley, H. *Nature Nanotechnol.* **2009**, *4*, 265–270.
- Howorka, S.; Cheley, S.; Bayley, H. *Nat. Biotechnol.* **2001**, *19*, 636–639.
- Ashkenasy, N.; Sánchez-Quesada, J.; Ghadiri, M.; Bayley, H. *Angew. Chem., Int. Ed.* **2005**, *44*, 1401–1404.
- Mitchell, N.; Howorka, S. *Angew. Chem., Int. Ed.* **2008**, *47*, 5565–5568.
- Wells, D.; Abramkina, A. *J. Chem. Phys.* **2007**, *127*, 125101–1–125101–10, A, A. .
- Phillips, J.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R.; Kale, L.; Schulten, K. *J. Comput. Chem.* **2005**, *26*, 1781–1802.
- Jarzynski, C. *Phys. Rev. Lett.* **1997**, *78*, 2690–2693.
- Aksimentiev, A.; Schulten, K. *Biophys. J.* **2005**, *88*, 3745–3761.
- Maglia, G.; Restrepo, M.; Mikhailova, E.; Bayley, H. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 19720–19725.
- Hornblower, B.; Coombs, A.; Whitaker, R.; Kolomeisky, A.; Picone, S.; Meller, A.; Akeson, M. *Nat. Methods* **2007**, 315–317.
- Kale, L.; Skeel, R.; Bhandarkar, M.; Brunner, R.; Gursoy, A.; Krawetz, N.; Phillips, J.; Shinozaki, A.; Varadarajan, K.; Schulten, K. *J. Comput. Phys.* **1999**, *151*, 283–312.
- Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graphics* **1996**, *14*, 33–38.
- Jorgensen, W.; Chandrasekhar, J.; Madura, J.; Impey, R.; Klein, M. *J. Chem. Phys.* **1983**, *79*, 926.
- Pearlman, D.; Case, D.; Caldwell, J.; Ross, W.; Cheatham III, T.; Debolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. *Comput. Phys. Commun.* **1995**, *91*, 1–41.
- Arnott, S.; Smith, P.; Chandrasekharan, R. *CRC Handbook of Biochemistry and Molecular Biology: Nucleic Acids*; CRC Press: Cleveland, OH, 1972.
- Wanunu, M.; Chakrabarti, B.; Mathé, J.; Nelson, D.; Meller, A. *Phys. Rev. E: Stat., Nonlinear, Soft Matter Phys.* **2008**, *77*, 31904.
- Wiggin, M.; Tropini, C.; Tabard-Cossa, V.; Jetha, N.; Marziali, A. *Biophys. J.* **2008**, *95*, 5317.
- Feller, S.; MacKerell, A. *J. Phys. Chem. B* **2000**, *104*, 7510–7515.
- Batcho, P. *J. Chem. Phys.* **2001**, *115*, 4003.
- Brünger, A. *X-Plor Version 3.1: A System for X-Ray Crystallography and NMR*; Yale University Press: New Haven, CT, 1993.
- Martyna, G.; Tobias, D.; Klein, M. *J. Chem. Phys.* **1994**, *101*, 4177.
- Park, S.; Khalili-Araghi, F. *J. Chem. Phys.* **2003**, *119*, 3559.
- Adler, A.; Grossman, L.; Fasman, G. *Proc. Natl. Acad. Sci. U.S.A.* **1967**, *57*, 423–430.
- Holcomb, D.; Tinoco Jr, I. *Pept. Sci.* **1965**, *3*, 121–133.
- Brahms, J.; Michelson, A.; Van Holde, K. *J. Mol. Biol.* **1966**, *15*, 467–88.
- Leng, M.; Felsenfeld, G. *J. Mol. Biol.* **1966**, *15*, 455–66.
- Jensen, M.; Park, S.; Tajkhorshid, E.; Schulten, K. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 6731.
- Ma, J.; Dougherty, D. *Chem. Rev.* **1997**, *97*, 1303–1324.
- Saksena, R.; Boghosian, B.; Fazendeiro, L.; Kenway, O.; Manos, S.; Mazzeo, M.; Sadiq, S.; Suter, J.; Wright, D.; Coveney, P. *Philos. Trans. R. Soc. London, Ser. A* **2009**, *367*, 2557–2571.

Modeling DNA Dynamics under Steady Deforming Forces and Torques

Alexey K. Mazur

CNRS UPR9080, Institut de Biologie Physico-Chimique, 13, rue Pierre et Marie Curie, Paris 75005, France

Received March 16, 2009

Abstract: An algorithm is developed for modeling atom-level dynamics of DNA subjected to steady external torques. For completeness, simulations with steady stretching loads are also considered. The algorithms were tested in Brownian dynamics simulations of discrete wormlike chain models with calibrated elastic properties to confirm that the elastic responses induced are of desired type and magnitude and that no side effects appear. The same methods were next used in a series of 100-ns all-atom MD simulations of tetradecamer DNA fragments with explicit water and counterions. The results demonstrate the possibility of probing regular elastic responses in DNA under low, nearly physiological amplitudes of forces and torques.

Introduction

Many fundamental biological processes are governed by the complex mechanics of the double-stranded DNA (dsDNA).¹ The dsDNA is very stiff compared to other polymers,^{2,3} and yet all its known X-ray structures are irregular and can be considered as deformed double helices.^{4,5} The deformations are attributed to specific base pair sequences, crystal packing forces, bound proteins, and so forth, indicating that the DNA double helix responds noticeably to any external perturbation, suggesting that this response is elastic. The conditions in the living cell are still more severe because the dsDNA is densely packed in chromosomes and forced to bend, twist, and stretch by numerous protein factors responsible for regulation of all vital processes.^{6,7} The stressed double helix, therefore, represents a major interest for experimental and theoretical investigations. In recent years, the nanomanipulation techniques made possible twisting and stretching of single dsDNA molecules using calibrated external forces and torques.⁸ These experiments provided rigorous tests of theories of polymer elasticity and further increased the theoretical interest of this field by revealing unforeseen phenomena like structural transitions under extreme stress and the coupling between small torsion and stretching deformations.^{9,10} Some results of these studies are controversial, notably, the nature of the overstretching transition^{11,12} and the measured value of the torsional stiffness.^{2,13}

The foregoing problems represent a major challenge for atom-level molecular modeling which, in principle, can trace

the effects of mesoscopic external stress in the detailed structure and dynamics of dsDNA. Already the pioneering computational studies of dsDNA probed its deformations using energy minimization in implicit solvent with bending potential restraints.¹⁴ Since then the power of computational methods gradually increased owing to improvement of force fields^{15–17} and simulation techniques.¹⁸ Modern free molecular dynamics (MD) simulations produce realistic conformational ensembles of dsDNA,¹⁹ which makes possible modeling of the stressed double helix with thermal fluctuations taken into account. In several recent investigations, restraining potentials were applied in the course of MD simulations to measure the free energies of strongly deformed states obtained by stretching,^{20,21} twisting,^{22,23} and bending²⁴ of dsDNA. Such use of conservative restraining forces follows the classical paradigm of molecular simulations,²⁵ but in some cases it would be advantageous to consider instead the DNA subjected to a steady stress. For instance, according to experiment, the torsional stiffness of circular dsDNA noticeably changes with small untwisting as a result of supercoiling.²⁶ One can use all-atom MD for measuring the torsional stiffness of short DNA,^{27,28} but to reproduce the conditions inside a supercoiled plasmid, a steady torque should be applied from both ends. Using potential restraints in this case is difficult, albeit possible. Importantly, simulations with steady forces and torques correspond to the most common experimental conditions *in vitro*; therefore, the computed thermodynamic averages would be directly com-

parable to experiment without differentiation of free energy profiles. An additional technical advantage is that one does not need to care about the absolute values of geometric parameters corresponding to different DNA lengths and sequences.

In the present article we describe and test an algorithm for modeling atom-level dynamics of dsDNA subjected to steady external torques. Simulations with steady stretching loads are considered as well, which is necessary for completeness because in experiments twisting and stretching are usually applied together. Steady stretching of dsDNA can also be modeled by using anisotropic pressure control.²⁹ To make sure that our algorithms induce elastic responses of desired type and magnitude and do not cause side effects, we first checked these methods in Brownian dynamics (BD) simulations of discrete wormlike chain (WLC) models of calibrated elasticity. The elastic parameters of DNA and the volume of sampling in these tests were chosen to be similar to those in atomistic simulations. These results were compared with two series of 100-ns all-atom MD simulations of tetradecamer DNA fragments with explicit water and counterions. Clear deviations from the WLC behavior were observed in MD in good agreement with earlier computations and experiments. Although very slow relaxation processes apparently take place in atomistic simulations, our results demonstrate the possibility of probing regular elastic responses with low, nearly physiological amplitudes of forces and torques.

Theory and Methods

Algorithms for Applying Steady Stress. The main idea of the algorithms described below is to load selected DNA bases with a set of external forces designed to cause the necessary deformation and then apply reactions to other bases to zero the total external force and torque. The technical difficulties are greatly reduced if the forces are applied to rigid Cartesian frames bound to the DNA structure, which is achieved by using virtual particles as shown in Figure 1. Four particles X_0 – X_3 are added to purine and pyrimidine bases as shown in Figure 1 for cytosine and guanine. A rigid Cartesian frame is centered at particle X_0 , with 1-Å distances X_0X_i ($i = 1, 2, 3$) and right angles between the corresponding vectors that define three orthogonal axes. The virtual particles are used only for applying external forces; they do not take part in other interactions. Some of them overlap with real atoms, others do not. Particle X_0 is placed close to the center of the canonical B-DNA, with vector X_0X_3 approximately parallel to the helical axis. To this end, X_1 coincides with N1 for purines and N3 for pyrimidines, vector X_0X_1 points along the exterior bisector of the corresponding planar angle, and the Cartesian frame is rotated around X_0X_1 to put X_2 in the plane of the base, with vector X_0X_3 pointing in a 5' to 3' direction in B-DNA.

Construction of dsDNA fragments is illustrated in Figure 2a. The modified bases are placed at the 3' ends of the two strands, with X_0X_3 vectors pointing in the opposite exterior directions. Modified versions of the four standard bases are needed for modeling arbitrary base pair sequences. In the

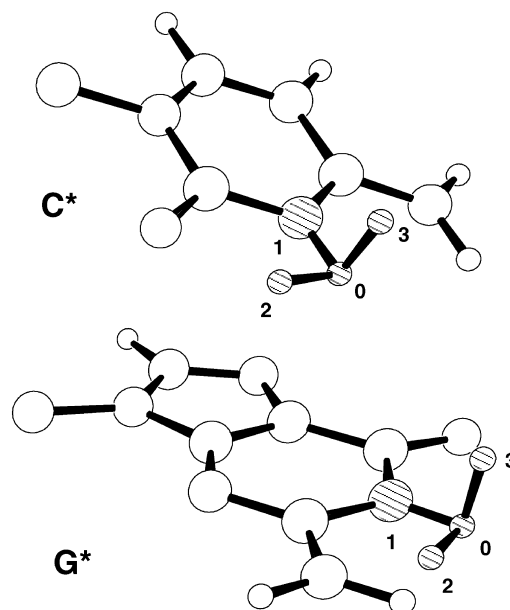


Figure 1. Construction of rigid local coordinate frames at nucleobases using virtual particles. Positions of the virtual particles are shown by dashed balls for cytosine and guanine. Similar modifications are applied to adenine and thymine. The perspective drawings were produced with ORTEP.⁵³

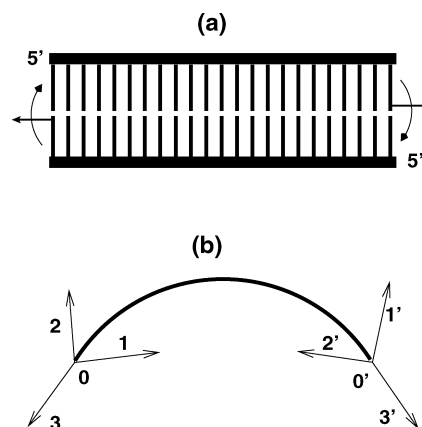


Figure 2. Construction of dsDNA with modified bases displayed in Figure 1. (a) A ladder schema of a dsDNA fragment of 23 bp. Modified bases are placed at the 3' ends of each strand. The arrows indicate the orientations of the terminal coordinate frames and the possible directions of applied torques. (b) The schema used for illustration of torques applied to curved DNA. The thick line represents the trace of a dsDNA fragment. The two terminal Cartesian frames are shown by vectors with the axes marked by the corresponding numbers of virtual particles. The right-hand frame is primed for distinction.

subsequent derivations we assume that the Cartesian frame is rigid and forms a united rigid body with the base. This is natural for the internal coordinate molecular dynamics (ICMD) method we use for DNA simulations.^{30,31} In the context of standard Newtonian dynamics these algorithms can be adapted with appropriate constraints and/or harmonic restraints applied to the virtual particles.

Consider a dsDNA fragment sketched in Figure 2b. We denote vectors X_iX_j and the corresponding unit vectors as \mathbf{r}_{ij} and \mathbf{e}_{ij} , respectively. Vectors \mathbf{e}_{03} and $\mathbf{e}_{0'3'}$ continue the

double helix in the two opposite directions. A torque \mathbf{T} around vector \mathbf{e}_{03} is created by a couple applied to particles X_0 and X_2 . A steady force of magnitude τ acts on X_0 along vector \mathbf{e}_{01} and an opposite force acts on X_2 . The resulting torque is $\mathbf{T} = \tau\mathbf{e}_{03}$. This vector can be decomposed into a component \mathbf{T}_{\parallel} parallel to vector $\mathbf{r}_{00'}$ and a perpendicular component denoted as \mathbf{T}_{\perp} . We have

$$\mathbf{T}_{\parallel} = \mathbf{e}_{00'}(\mathbf{T} \cdot \mathbf{e}_{00'}) = \tau\mathbf{e}_{00'}(\mathbf{e}_{03} \cdot \mathbf{e}_{00'})$$

and

$$\mathbf{T}_{\perp} = \mathbf{T} - \mathbf{T}_{\parallel}$$

Component \mathbf{T}_{\perp} is compensated by a reaction couple applied to particles X_0 and X'_0 . Its value $f_{00'}$ and direction \mathbf{e}_{\perp} are obtained as

$$f_{00'} = \tau\mathbf{e}_{03} - \mathbf{e}_{00'}(\mathbf{e}_{03} \cdot \mathbf{e}_{00'})/|\mathbf{r}_{00'}|$$

and

$$\mathbf{e}_{\perp} = \pm(\mathbf{e}_{03} \times \mathbf{e}_{00'})/|\mathbf{e}_{03} \times \mathbf{e}_{00'}|$$

When vectors \mathbf{e}_{03} and $\mathbf{e}_{00'}$ are collinear, these reactions are unnecessary and $f_{00'} = 0$.

Component \mathbf{T}_{\parallel} is projected onto vectors $\mathbf{e}_{01'}$, $\mathbf{e}_{02'}$, and $\mathbf{e}_{03'}$ and compensated by three reaction couples applied to particles $X_{0'}-X_{3'}$. For instance, the component along vector $\mathbf{e}_{03'}$ can be compensated by forces applied to $X_{0'}$ and $X_{2'}$ parallel to $\mathbf{e}_{01'}$ with magnitudes

$$f_{02'} = \pm(\mathbf{e}_{03'} \cdot \mathbf{T}_{\parallel})$$

The other two reaction couples are computed similarly. One can choose between several alternative ways of applying these reactions, which does not affect the results. In principle, the original torque \mathbf{T} could be directly projected onto the opposite coordinate frame and compensated by reaction couples applied to its axes. With some orientations, however, these reactions would correspond to bending rather than twisting. The decomposition into \mathbf{T}_{\parallel} and \mathbf{T}_{\perp} reduces the magnitudes of undesirable nontwisting reactions applied to the molecule.

By construction, the foregoing operations result in zero total force and torque on the DNA fragment. Since the individual forces are applied at different points, internal stress and deformations are introduced that should correspond to overall twisting. It is instructive to consider some extreme cases. If \mathbf{e}_{03} and $\mathbf{e}_{03'}$ are antiparallel as in straight or circular DNA, $\mathbf{T}_{\perp} = 0$, the original couple and its torque are compensated by $f_{02'} = \tau$, and we obtain pure torsional stress. In contrast, when \mathbf{e}_{03} and $\mathbf{e}_{03'}$ are parallel as in a U-turn, the original torque is completely compensated by $f_{00'}$, with no reactions applied to the primed coordinate frame. In this case rotation of the dsDNA fragment around its curved axis does not affect the total angular momentum and becomes a kind of internal degree of freedom. The external torque appears to be opposed only by solvent friction; therefore, the kinetic energy would be pumped into the system, leading to a temperature explosion. To reduce the probability of such situations, the external stress is applied symmetrically; that

is, two torques of 0.5τ are applied from the opposite DNA ends and compensated as described above.

The same terminal coordinate frames were used for other types of external stress. A steady stretching load was created by two opposite forces of magnitude f along vector $X_0X_{0'}$. Bending stress can be obtained similarly to twisting by changing the direction of the original torque \mathbf{T} so that the corresponding couple would be applied to particles X_0 and X_3 in a locally fixed plane. The case of bending strain was not studied here because it is mainly relevant to protein–DNA interactions, which involves essential complications that go beyond the present study.

Brownian Dynamics. The behavior of dsDNA consistent with the WLC theory^{3,32} was simulated by Brownian dynamics of a composite-bead chain described in a recent report.²⁸ Every base pair was represented by a rigid composite bead of four virtual particles labeled $P_0^j-P_3^j$ that form a rigid Cartesian frame similar to those in Figure 1. The upper index refers to the consecutive bead number in the chain. The total energy of an N -bead chain is composed of bond stretching, torsion, and bending terms as

$$U = \frac{h}{2} \sum_{i=1}^{N-1} (l_i - l_0)^2 + \frac{q}{2} \sum_{i=1}^{N-1} (\phi_i - \phi_0)^2 + \frac{g}{2} \sum_{i=1}^{N-2} \theta_i^2 \quad (1)$$

All other interactions between the beads can be neglected because we consider only short stiff chains that cannot loop. Here, l_i is the distance between particles P_0^i and P_0^{i+1} and l_0 is its equilibrium value. The l_0 value was 0.35 nm, corresponding to one base pair step. The torsion potential was applied to dihedral angles $\phi_i = P_1^i P_0^i P_0^{i+1} P_1^{i+1}$, with the equilibrium helical twist angle $\phi_0 = 0$ used for convenience. In the bend energy term, the angular displacement was measured for angles $\theta_i = P_0^i P_0^{i+1} P_0^{i+2}$.

The DNA elasticity is conveniently characterized by three persistence lengths (PLs) corresponding to bending, twisting, and stretching that we denote here as l_b , l_t , and l_s , respectively. Parameters g , q , and h in eq 1 were chosen so that the PL values were similar in BD and MD. The energy coefficients were first estimated analytically and then verified by processing the resulting BD trajectories with previously described procedures.^{28,33} For bending and twisting we applied $l_b = 73$ nm and $l_t = 137$ nm. The stretching PL is difficult to measure accurately.²⁸ Here, we used $l_s = 100$ nm corresponding to the Young's modulus $Y_f = 1410$ pN. The average DNA length and the overall winding were evaluated from the corresponding instantaneous values as $\langle L = \sum_{i=1}^{N-1} l_i \rangle$ and $\langle \Phi = \sum_{i=1}^{N-1} \phi_i \rangle$, respectively, where the angular brackets denote time averaging. These values are appropriate for discussing macroscopic elastic responses. Note also that the fluctuations of $\langle L \rangle$ and $\langle \Phi \rangle$ in real DNA cannot be obtained by averaging base-step parameters because of the near-neighbor correlations. The bending is conveniently characterized by parameter $\langle 1 - \cos(\Theta) \rangle$ where angle Θ was measured between vectors X_0X_3 and $X_{0'}X_{3'}$.

The BD simulation algorithm was considered in detail in a previous report.²⁸ It is based on earlier results of different groups.^{34–38} For better comparison with atom-level MD, we

considered short chains of 14 composite beads, computed BD trajectories of 100 ns with a time step of 0.005 ps, and saved chain configurations every 5 ps. The reduced time step removed a small discrepancy between the measured and the analytical values of the torsional PL and improved the even distribution of torsional fluctuations within the chains.

Molecular Dynamics. The classical MD simulations were carried out using previously described protocols.^{27,28} The starting states were prepared as follows. A tetradecamer DNA fragment with the canonical B-DNA conformation³⁹ was immersed in a 6.2-nm cubic cell with a high water density of 1.04. The box was neutralized by placing Na⁺ ions at random water positions at least 5 Å from the solute. The system was energy minimized and dynamics were initiated with the Maxwell distribution of generalized momenta at low temperature. The system was next slowly heated to 293 K and equilibrated for 1.0 ns. After that, the water density was adjusted to 0.997 by removing the necessary number of water molecules selected randomly at least 5 Å from DNA and ions, and the simulations were continued with NVT ensemble conditions. The temperature was maintained by the Berendsen algorithm⁴⁰ applied separately to DNA, water, and ions, with a relaxation time of 10 ps. Simulations with external forces and torques started from equilibrated states after a few nanoseconds of free dynamics.

The AMBER98 force field parameters^{15,16} were used with the rigid TIP3P water model.⁴¹ The electrostatic interactions were treated by the SPME method,¹⁸ with the common values of Ewald parameters, that is, 9-Å truncation for the real space sum and $\beta \approx 0.35$. To increase the time step, MD simulations were carried out by the ICMD method,^{42,43} with the internal DNA mobility limited to essential degrees of freedom and rotation of water molecules and internal DNA groups including only hydrogen atoms slowed down by weighting of the corresponding inertia tensors.^{44,45} The double-helical DNA was modeled with all backbone torsions, free bond angles in the sugar rings, and rigid bases and phosphate groups. The effect of these constraints is insignificant, as was previously checked through comparisons with standard Cartesian dynamics.^{27,44} The time step was 0.01 ps and the DNA structures were saved every 5 ps. All trajectories were continued to 100 ns and the sampled conformations of the double helix were analyzed by program 3DNA.⁴⁶

In all MD simulations we used the tetradecamer AT-alternating dsDNA fragment (*d(AT)₇*). This choice is consistent with recent computations^{28,33,33} and it was dictated by the following considerations. The length, slightly larger than one helical turn, is convenient for measuring the elastic parameters of DNA.²⁸ These molecules are homopolymers of AT units; therefore, they cannot have distinguished asymmetric structures like static bends. True homopolymer DNA duplexes have special features and, in free MD with the AMBER force field, these structures deviate from the canonical B-DNA, stronger than AT- and GC-alternating sequences.⁴⁷ The terminal AT base pairs open rather frequently during nanosecond time scale MD, which significantly perturbs the flanking DNA structure. Because this dynamics cannot be averaged during the accessible duration

of MD trajectories, we blocked it by applying upper distance restraints similar to those used in NMR-based structure calculations.⁴⁸ The restraints of the following form

$$U_r(r_{ij}) = \begin{cases} A_r(r_{ij} - r_0)^2, & r_{ij} > r_0 \\ 0, & r_{ij} \leq r_0 \end{cases}$$

were applied to atoms of Watson–Crick hydrogen bonds in terminal base pairs. The cutoff r_0 was 3.6 and 2.5 Å, respectively, for the donor–acceptor and hydrogen–acceptor distances. The value of A_r was 1 kcal·mol·Å⁻² for both. Because of the relatively large r_0 distances, these restraints operate only during rare and short intervals when the base pairing is close to rupture. Analysis of trajectories showed that the probability of such events varied from 0.03 to 3.5% for different restraints, and that the fraction of states with nonzero restraint energy was below 7.4%. Therefore, the spurious perturbations of averages can be considered negligible.

The MD simulations with the AMBER force field were earlier shown to suffer from rare irreversible transitions of backbone torsion angles to nonstandard states, which was attributed to force field artifacts.^{17,49} Analysis of our trajectories revealed a pattern of conformational substates qualitatively similar to those found in earlier studies, but the relative populations of noncanonical conformers were not as large as those in the report by Dixit et al.⁴⁹ The irreversible $\gamma^{s+} \rightarrow \gamma'$ transitions accompanied by α^{s-}/α^{s+} and β'/β^{s+} dynamics were observed in all the terminal ApT steps. Therefore, only the central dodecamer was chosen for the analysis. A few $\gamma^{s+} \rightarrow \gamma'$ transitions within this fragment are mentioned in the Results section. Their effect was not essential for the purposes of the present study.

Results

Validation by Brownian Dynamics. The artificial external stress is meant to be applied in the course of long all-atom MD simulations, with the DNA chain taking many different conformations. Correct algorithms should produce perturbations of desired type and magnitude and should not give side effects like compressing or stretching along with twisting, for instance. It is essential to check for such artifacts before application in MD; otherwise, small artificial correlations between twisting and stretching can be confused with similar properties of real DNA observed in experiments.^{9,10} The BD simulations of discrete WLC models of dsDNA are well-suited for the verification tests because they correctly sample from canonical ensembles of chain configurations and also give an appreciation of convergence to be expected with durations of trajectories accessible in MD.^{28,33}

The effect of the steady torsional stress on the BD of the discrete WLC model is shown in Figures 3 and 4. The DNA length is just a little longer than one helical turn, which is appropriate for probing the DNA elasticity by all-atom MD simulations.²⁸ Accordingly, the duration of trajectories and the elastic parameters of DNA correspond to the common MD conditions. The range of torque values covers that used in single-molecule experiments¹³ as well as the range of stability of real DNA. The BD results confirm that the algorithm correctly works in these conditions and that it can

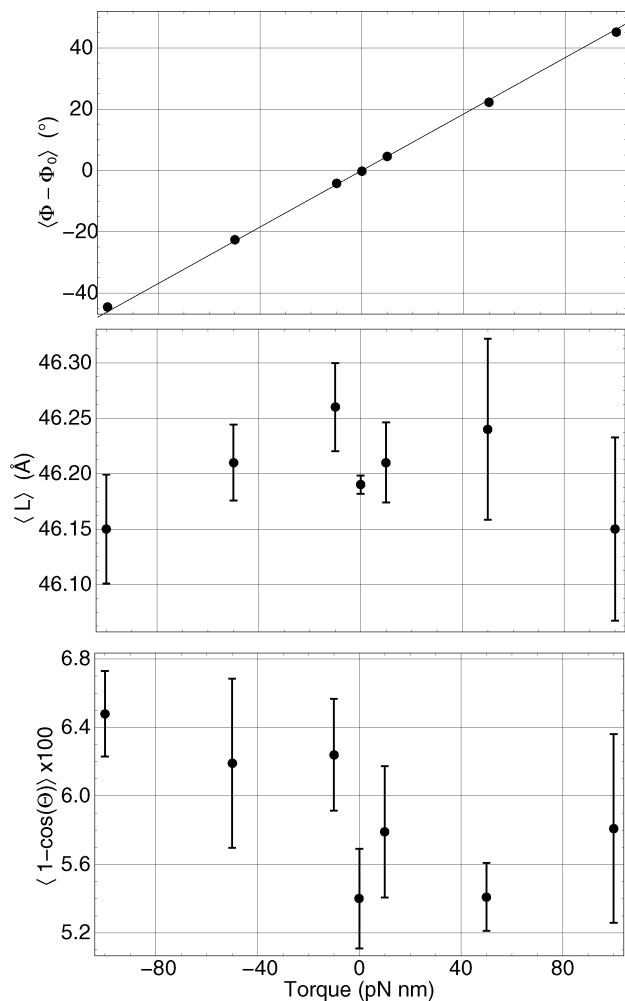


Figure 3. BD simulations of DNA with applied external torques of different magnitudes. The results are shown for the overall twisting (top), DNA length (middle), and bending (bottom). The straight line on the top panel shows the theoretical dependence for $l_t = 137$ nm. The error bars show statistical errors evaluated by the method of block averages.²⁵ Every trajectory was divided into four equal parts (blocks) and the errors are estimated from standard deviations of the block averages. For twisting the errors are smaller than the size of the plot points.

hardly cause spurious artifacts in MD. The overall DNA winding changes linearly with the torque, in quantitative agreement with theory. The torsional stress is evenly distributed over the whole chain (see Figure 4). The DNA length is constant within the error limits. The stretching/compression side effects are negligible and indistinguishable by practical methods of measuring the length of all-atom DNA. In long DNA, the external torsional stress causes buckling transitions and here it should increase bending. Figure 3 does not contradict that, but for one DNA turn this effect is small.

Similar tests for steady stretching are shown in Figures 5 and 6. In single-molecule experiments forces up to 45 pN are used to keep DNA straight¹³ and the B-DNA double helix remains stable with forces below 70 pN. It can be seen that within this range of forces the steady stretching algorithm does not cause spurious side effects. The DNA length increases in quantitative agreement with theory. The stretch-

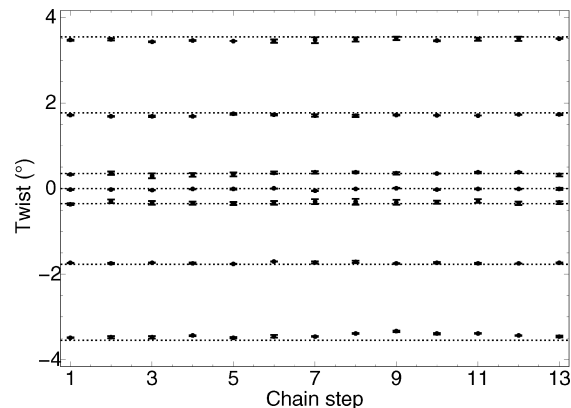


Figure 4. BD simulations of DNA with applied external torques of different magnitudes. The distribution of the torsional deformations along the chain. The torque values are the same as those in Figure 3. Theoretical values corresponding to a WLC model with $l_t = 137$ nm are shown by horizontal dotted lines. The errors are similar to the size of the plot points.

ing stress is evenly distributed along the chain (see Figure 6). Figure 6 looks more noisy than Figure 4, even though the respective ranges of stretching and torsion deformations correspond to similar energies. This difference is due to very fast single-step torsion relaxation²⁸ that results in a better convergence for 100 ns of BD. The overall winding fluctuates near zero without distinguishable trends. The bending steadily decreases. Interestingly, it is reduced by less than two times and does not level even with very large forces, which means that DNA stretching always leaves some freedom for local bending.

Molecular Dynamics. The results of the MD tests are displayed in Figures 7–10. The data analysis involved the central part of a tetradecamer AT-alternating dsDNA. The choice of this specific fragment was explained in the Theory and Methods section. Regular elastic responses to twisting and stretching are shown, respectively, in the top panel of Figure 7 and the middle panel of Figure 9. The overall winding Φ exhibits linear variations with torque, in good qualitative and quantitative agreement with the WLC theory, as well as the earlier measured twisting stiffness. The $\langle L(f) \rangle$ dependence is clearly ascending as it should be, but the statistical errors are larger than those in the BD tests in Figure 5 and one can only say that the amplitude of the extension roughly corresponds to the stretching Young's modulus of 1100 pN. This parameter is intrinsically difficult to measure accurately in MD simulations of short DNA.²⁸

Compared to the corresponding BD plots, MD exhibits an evidently broader scattering of points around the theoretical straight lines, indicating the presence of slow processes with relaxation times beyond those observed with the discrete WLC models. Very slow processes of still unclear nature were experimentally detected in short dsDNA.^{50,51} However, a few systematic shifts of the plot points in Figures 7 and 9 correlate with, and are probably caused by, the rare $\gamma^{g+} \rightarrow \gamma'$ backbone transitions (see the Theory and Methods section). An isolated transition such as this does not cause gross changes in the DNA structure, but it results in a stepwise increase of the local twist and rise and, accordingly, augments

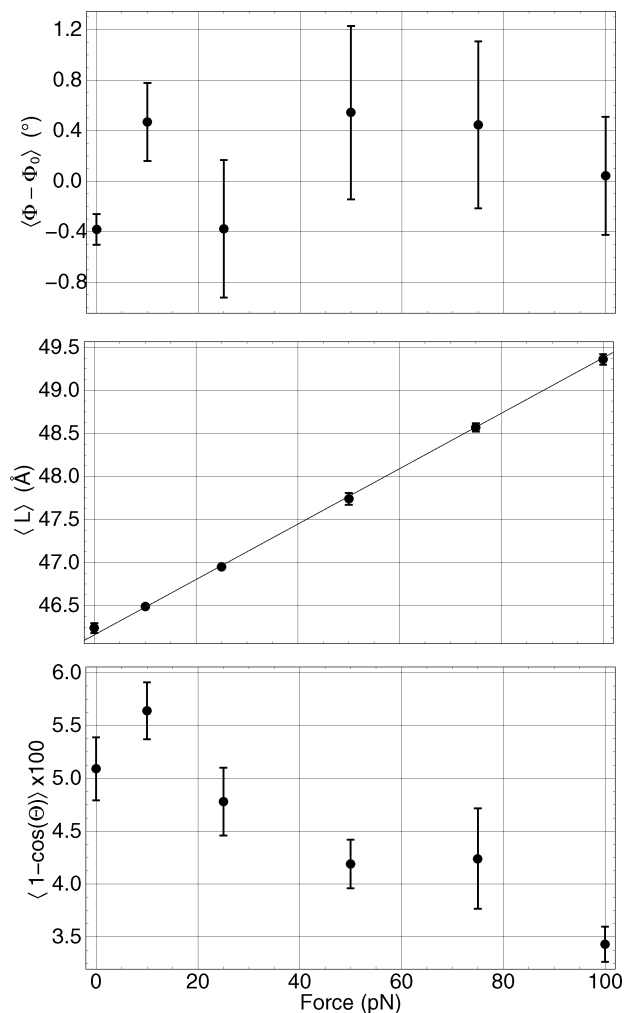


Figure 5. BD simulations of DNA with applied external stretching forces of different magnitudes. The results are shown for the overall twisting (top), DNA length (middle), and bending (bottom). The straight line on the middle panel shows the theoretical dependence for $l_s = 100$ nm. In this panel, the statistical errors are similar to the size of the plot points.

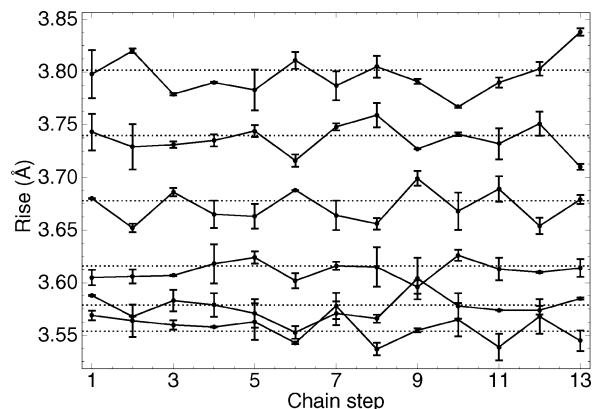


Figure 6. BD simulations of DNA with applied external stretching force of different magnitudes. The distribution of the stretching deformations along the chain. The force values are the same as those in Figure 5. The horizontal dotted lines mark the theoretical levels corresponding to $l_s = 100$ nm.

the measured overall DNA length and winding. In Figure 7, for instance, no $\gamma^{g+} \rightarrow \gamma^f$ transitions occurred with zero and

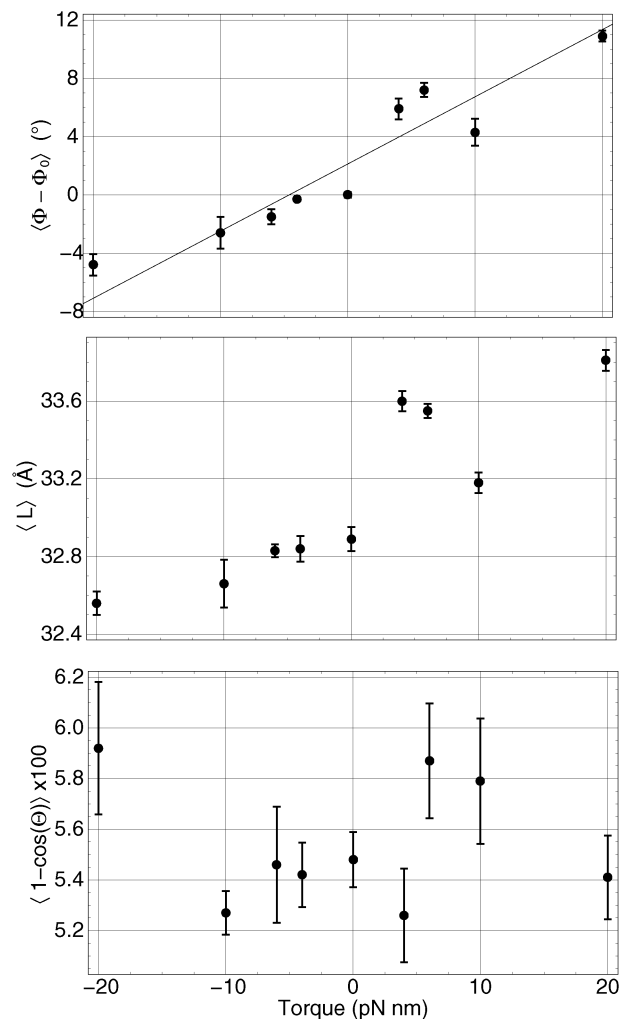


Figure 7. MD simulations of DNA under external twisting torque of different magnitudes. The results are shown for the overall twisting (top), DNA length (middle), and bending (bottom). The straight line on the top panel shows the theoretical dependence for $l_t = 137$ nm.

negative torques. With $\tau = 10$ pN two isolated transitions occurred, but they reversed after about 30 and 60 ns, respectively. Finally, in each of the trajectories with $\tau = 4, 6,$ and 20 pN one nonreversed transition took place. These rare events evidently correlate with the small shifts of the plot points in Figure 7. The stretching data were affected by three isolated nonreversed $\gamma^{g+} \rightarrow \gamma^f$ transitions. One of them occurred with $f = 10$ pN, but it happened late in the dynamics and did not cause noticeable effects. The other two transitions took place during dynamics under 4 pN stretching, and the corresponding data points are clear outliers (see Figure 9).

The above rare conformational transitions do not hide the regular trends produced by the applied external stress. The middle panel of Figure 7 shows the chiral effect revealed in single-molecule experiments,^{9,10} namely, that the DNA length is increased with overtwisting and reduced with unwinding. The pattern in Figure 7 is qualitatively different from that in Figure 3, which proves that this subtle feature is not a mere artifact. With a torque of 20 pNnm, the increase in the DNA length extrapolated from experimental data⁹ should be around 0.14 \AA , which agrees reasonably well with Figure 7. No statistically significant trends could be detected

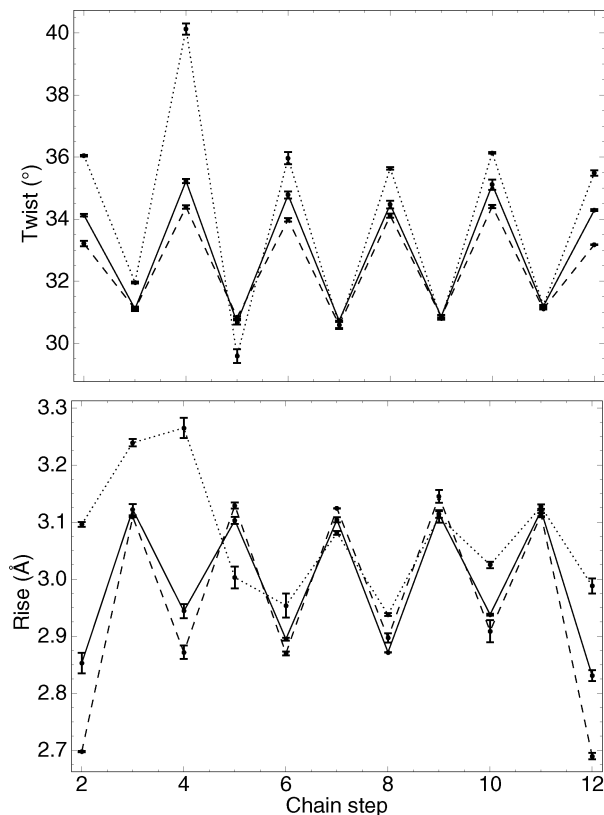


Figure 8. MD simulations of DNA under external twisting torque of different magnitudes. Distribution of perturbations along the chain for twisting (upper panel) and stretching (lower panel). The torque values are 0 pN·nm (solid line), -20 pN·nm (dashed line), and $+20$ pN·nm (dotted line).

for bending, which qualitatively agrees with the BD data (compare the bottom panels of Figures 7 and 3).

Figure 8 demonstrates that the external torque is evenly distributed over the double helix. The ApT and TpA steps are characterized by strongly different values of equilibrium intrinsic twist and rise. The figure shows that the TpA steps are much more receptive to forced twisting than the ApT steps. This agrees with earlier experimental and computational observations.^{5,22} In simulations with the overtwisting torque of 20 pN·nm the fourth TpA step exhibited anomalous behavior because of a $\gamma^{s+} \rightarrow \gamma'$ transition in one of the strands. This transition changes the equilibrium intrinsic twist, which corresponds to plastic (nonelastic) deformability, but it can also affect the local stiffness. The noncanonical α/γ backbone conformations are considered as force field artifacts in free MD simulations, but such conformers are frequent in protein–DNA complexes.⁵² The anomalous stiffness of such conformers may be important for biology and this issue deserves future study. The lower panel in Figure 8 demonstrates also that the lengths of the terminal base pair steps are affected by external twisting stronger than others.

The results of MD simulations with steady external stretching are analyzed in Figures 9 and 10. With the stretching load most common for experiments ($f < 20$ pN)^{9,13} only small perturbations of the DNA structure and dynamics are distinguishable. The chiral twist–stretch coupling revealed in the middle panel of Figure 7 should result in

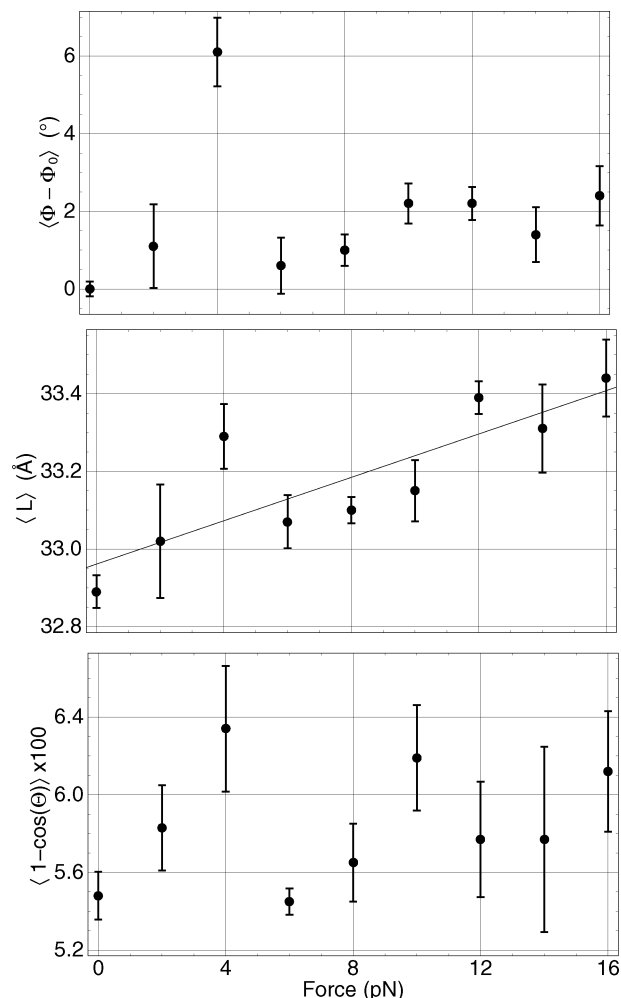


Figure 9. MD simulations of DNA under external stretching force of different magnitudes. The results are shown for the overall twisting (top), DNA length (middle), and bending (bottom). The straight line on the middle panel shows the best fit theoretical dependence corresponding to $l_s = 83$ nm ($Y_f = 1170$ pN).

overwinding with increased force. According to experiment,⁹ for the stretching load of 16 pN, the $\Delta\Phi$ value should reach ca. 0.6° . Our results do not contradict this estimate (cf. top panel of Figure 9), but the effect is within the amplitude of statistical fluctuations and smaller than the shift at 4 pN caused by two $\gamma^{s+} \rightarrow \gamma'$ transitions mentioned above. It might be generally difficult to detect such small change with the DNA length accessible in MD because the sign of the coupling is apparently inverted with forces beyond 20 pN.⁹ The bottom panel of Figure 9 exhibits bending fluctuations comparable with the corresponding plate in Figure 5 for BD. The perturbations produced by external stretching are evenly distributed along the double helix (Figure 10). Similarly to Figure 8 the TpA steps are noticeably more susceptible to deformations than the ApT steps. In contrast, the terminal base pair steps do not show excessive deformability as those shown in Figure 8; moreover, they seem to be more rigid than the internal steps.

Discussion

This article presents a nonstandard approach to modeling deformed states of dsDNA. Replacing potential geometric

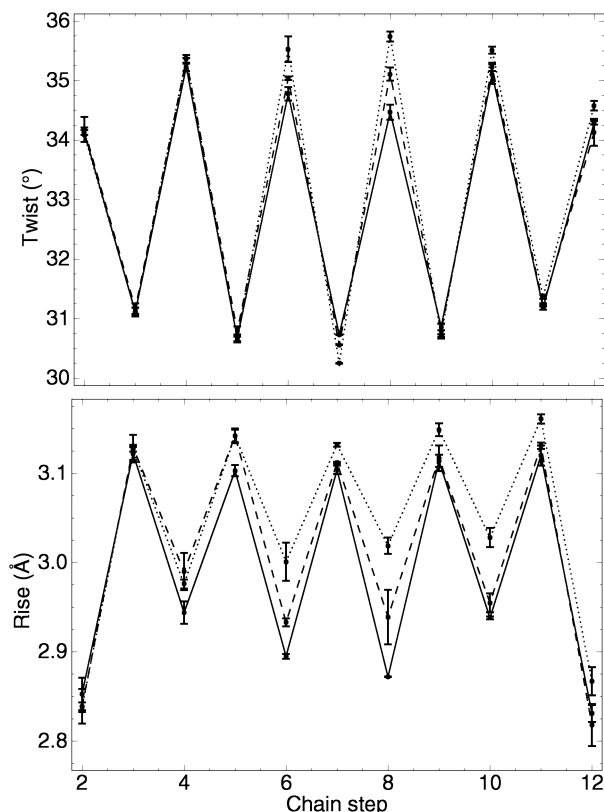


Figure 10. MD simulations of DNA under external stretching force of different magnitudes. Distribution of perturbations along the chain for twisting (upper panel) and stretching (lower panel). The force values are 0 pN·nm (solid line), 8 pN (dashed line), and 16 pN (dotted line).

restraints by nonpotential forces and torques is analogous to the switch from constant volume to constant pressure simulations of continuous media. The steady stress conditions correspond to the state of short dsDNA fragments in living cells and also in single-molecule experiments. To my knowledge, the present study is the first such attempt. Methods of this type will serve as a useful supplement to the earlier setups using potential restraints and umbrella sampling.^{20–22,24}

Steady forces are seldom used in MD simulations because it is difficult to anticipate situations where the system appears outside domains of attraction and dynamics become unbound. One relevant example was considered above in the derivation of the steady torque algorithm. The possibility of other potentially dangerous configurations cannot be excluded for long dsDNA. Our tests demonstrate that the algorithms proposed are sufficiently robust for small DNA lengths and the values of torques and forces corresponding to experimental and physiological conditions. I believe, however, that the same approach can be safely applied to somewhat longer dsDNA, notably, that it can be used for modeling the buckling transition in medium length fragments. Such studies are planned.

The results of MD simulations reported here are preliminary. They confirm the correctness of the algorithms and the principal possibility of imitating the steady stress conditions in all-atom MD of dsDNA. These computations are continued to obtain a more representative sampling and

to check the possible effect of the external stress on the elastic parameters of dsDNA. Additional longer simulations are planned to probe such effects for other biologically relevant base pair sequences. Despite the above limitations, the results reported here illustrate the power of modern MD methods; notably, we demonstrated that atomistic simulations can reproduce linear elastic response of dsDNA in quantitative agreement with the WLC theory and the elastic parameters previously evaluated by other methods. The present data also give an appreciation of the complexity and the richness of real DNA dynamics compared to the discrete WLC models and BD simulations.

An important technical problem in long-time MD simulations of dsDNA is related to rare noncanonical backbone transitions.⁴⁹ The number of such events here was less significant than might be expected, and moreover, in two cases reverse transitions occurred. This can be attributed to (i) the use of AMBER98 rather than other similar force field versions and (ii) a fortunate choice of the base pair sequence. Free AMBER simulations are known to underestimate the helical twist in dsDNA, but this effect is less pronounced for some sequences and it was specifically reduced by the AMBER98 force field modifications.¹⁶ The overall shift of the dsDNA structure toward the A-form may favor the noncanonical backbone transitions. Future simulations should use the recently improved force field version¹⁷ to reduce the probability of such events. However, the noncanonical backbone conformations are observed in protein–DNA complexes;⁵² therefore, the properties of these conformers should be studied as well.

Traditional approaches using potential restraints and umbrella sampling are designed to improve the sampling of weakly populated states; therefore, they facilitate probing strong DNA deformations. For small perturbations considered here these methods are not efficient because the shifts in the restrained values would be smaller than the natural thermal fluctuations. In such cases, the potentials of mean force (PMF) can be obtained from a single unrestrained computation. However, very high accuracy of the PMF profile is necessary for extracting the first and second derivatives required for the evaluation of the elastic parameters of DNA and comparisons with experiment. In the published umbrella sampling calculations^{20–22} the PMFs obtained were below the noise level in the range of forces and torques considered here. At the same time, a number of intriguing features in the biological function of dsDNA involve very small changes in the helical twist, for instance. The steady stress algorithms are specifically designed for probing such conditions; therefore, they should find many interesting applications.

References

- (1) Bloomfield, V. A.; Crothers, D. M.; Tinoco, I. *Nucleic Acids. Structures Properties and Functions*; University Science Books: Sausalito, CA, 2000.
- (2) Hagerman, P. J. *Annu. Rev. Biophys. Biophys. Chem.* **1988**, *17*, 265–286.
- (3) Cantor, C. R.; Schimmel, P. R. *Biophysical Chemistry, Part III: The Behavior of Biological Macromolecules*; W.H. Freeman: San Francisco, 1980.

- (4) Dickerson, R. E. In *Oxford Handbook of Nucleic Acid Structure*; Neidle, S., Ed.; Oxford University Press: New York, 1999; pp 145–197.
- (5) Olson, W. K.; Gorin, A. A.; Lu, X. J.; Hock, L. M.; Zhurkin, V. B. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 11163–11168.
- (6) Widom, J. *Q. Rev. Biophys.* **2001**, *34*, 269–324.
- (7) Cozzarelli, N. R.; Cost, G. J.; Nollmann, M.; Viard, T.; Stray, J. E. *Nat. Rev. Mol. Cell. Biol.* **2006**, *7*, 580–588.
- (8) Bustamante, C.; Bryant, Z.; Smith, S. B. *Nature* **2003**, *421*, 423–427.
- (9) Gore, J.; Bryant, Z.; Nollmann, M.; Le, M. U.; Cozzarelli, N. R.; Bustamante, C. *Nature* **2006**, *442*, 836–839.
- (10) Lionnet, T.; Joubaud, S.; Lavery, R.; Bensimon, D.; Croquette, V. *Phys. Rev. Lett.* **2006**, *96*, 178102.
- (11) Cluzel, P.; Lebrun, A.; Heller, C.; Lavery, R.; Viovy, J. L.; Chatenay, D.; Caron, F. *Science* **1996**, *271*, 792–794.
- (12) Rouzina, I.; Bloomfield, V. A. *Biophys. J.* **2001**, *80*, 882–893.
- (13) Bryant, Z.; Stone, M. D.; Gore, J.; Smith, S. B.; Cozzarelli, N. R.; Bustamante, C. *Nature* **2003**, *424*, 338–341.
- (14) Zhurkin, V. B.; Lysov, Y. P.; Ivanov, V. I. *Nucleic Acids Res.* **1979**, *6*, 1081–1096.
- (15) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
- (16) Cheatham, T. E., III; Cieplak, P.; Kollman, P. A. *J. Biomol. Struct. Dyn.* **1999**, *16*, 845–862.
- (17) Perez, A.; Marchan, I.; Svozil, D.; Sponer, J.; Cheatham, T. E.; Laughton, C. A.; Orozco, M. *Biophys. J.* **2007**, *92*, 3817–3829.
- (18) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.
- (19) Cheatham, T. E., III; Kollman, P. A. *Annu. Rev. Phys. Chem.* **2000**, *51*, 435–471.
- (20) MacKerell, A. D., Jr.; Lee, G. U. *Eur. Biophys. J.* **1999**, *28*, 415–426.
- (21) Harris, S. A.; Sands, Z. A.; Laughton, C. A. *Biophys. J.* **2005**, *88*, 1684–1691.
- (22) Kannan, S.; Kohlhoff, K.; Zacharias, M. *Biophys. J.* **2006**, *91*, 2956–2965.
- (23) Wereszczynski, J.; Andricioaei, I. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 16200–16205.
- (24) Curuksu, J.; Zakrzewska, K.; Zacharias, M. *Nucleic Acids Res.* **2008**, *36*, 2268–2283.
- (25) Frenkel, D.; Smit, B. *Understanding Molecular Simulations. From Algorithms to Applications*; Academic Press: New York, 1996.
- (26) Selvin, P. R.; Cook, D. N.; Pon, N. G.; Bauer, W. R.; Klein, M. P.; Hearst, J. E. *Science* **1992**, *255*, 82–85.
- (27) Mazur, A. K. *Biophys. J.* **2006**, *91*, 4507–4518.
- (28) Mazur, A. K. *J. Phys. Chem. B* **2009**, *113*, 2077–2089.
- (29) Luan, B.; Aksimentiev, A. *Phys. Rev. Lett.* **2008**, *101*, 118101.
- (30) Mazur, A. K. *J. Am. Chem. Soc.* **2000**, *122*, 12778–12785.
- (31) Mazur, A. K. *J. Am. Chem. Soc.* **2002**, *124*, 14707–14715.
- (32) Landau, L. D.; Lifshitz, E. M. *Statistical Physics, Part I*; Nauka: Moscow, 1976.
- (33) Mazur, A. K. *J. Phys. Chem. B* **2008**, *112*, 4975–4982.
- (34) Ermak, D. L.; McCammon, J. A. *J. Chem. Phys.* **1978**, *69*, 1352–1360.
- (35) Allison, S. A. *Macromolecules* **1986**, *19*, 118–124.
- (36) Iniesta, A.; de la Torre, J. G. *J. Chem. Phys.* **1990**, *92*, 2015–2018.
- (37) Chirico, G.; Langowski, J. *Macromolecules* **1992**, *25*, 769–775.
- (38) Jian, H.; Schlick, T.; Vologodskii, A. *J. Mol. Biol.* **1998**, *284*, 287–296.
- (39) Arnott, S.; Hukins, D. W. L. *Biochem. Biophys. Res. Commun.* **1972**, *47*, 1504–1509.
- (40) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (41) Jorgensen, W. L.; Chandreskhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (42) Mazur, A. K. *J. Comput. Chem.* **1997**, *18*, 1354–1364.
- (43) Mazur, A. K. *J. Chem. Phys.* **1999**, *111*, 1407–1414.
- (44) Mazur, A. K. *J. Am. Chem. Soc.* **1998**, *120*, 10928–10937.
- (45) Mazur, A. K. *J. Phys. Chem. B* **1998**, *102*, 473–479.
- (46) Lu, X.-J.; Olson, W. K. *Nucleic Acids Res.* **2003**, *31*, 5108–5121.
- (47) Mazur, A. K. *J. Chem. Theory Comput.* **2005**, *1*, 325–336.
- (48) Kozin, S. A.; Bertho, G.; Mazur, A. K.; Rabesona, H.; Girault, J. P.; Haertle, T.; Takahashi, M.; Debey, P.; Hoa, G. H. *J. Biol. Chem.* **2001**, *276*, 46364–46370.
- (49) Dixit, S. B.; Beveridge, D. L.; Case, D. A.; Cheatham, T. E., III; Giudice, E.; Lankas, F.; Lavery, R.; Maddocks, J. H.; Osman, R.; Sklenar, H.; Thayer, K. M.; Varnai, P. *Biophys. J.* **2005**, *89*, 3721–3740.
- (50) Brauns, E. B.; Madaras, M. L.; Coleman, R. S.; Murphy, C. J.; Berg, M. A. *Phys. Rev. Lett.* **2002**, *88*, 158101.
- (51) Sen, S.; Andreatta, D.; Ponomarev, S. Y.; Beveridge, D. L.; Berg, M. A. *J. Am. Chem. Soc.* **2009**, *131*, 1724–1735.
- (52) Varnai, P.; Djuranovic, D.; Lavery, R.; Hartmann, B. *Nucleic Acids Res.* **2002**, *30*, 5398–5406.
- (53) Burnett, M. N.; Johnson, C. K. *ORTEP-III: Oak Ridge Thermal Ellipsoid Plot program for crystal structure illustrations*; Oak Ridge National Laboratory Report ORNL-6895; Oak Ridge National Laboratory: Oak Ridge, TN, 1996. CT900124N

Ab Initio Raman Spectra of β -Lactamase Inhibitor Intermediates Bound to E166A SHV β -Lactamase

Andrea Miani,[†] Marion Skalweit Helfand,[‡] and Simone Raugei^{*†}

SISSA and INFN-DEMOCRITOS center via Beirut 2, I-34014, Trieste, Italy, and Louis Stokes Cleveland Department of Veterans Affairs Medical Center, Cleveland, Ohio 44106

Received March 20, 2009

Abstract: The assignment and the analysis of the experimental vibrational Raman spectra of enzyme bound β -lactamase inhibitors may be of help to understand the mechanisms responsible for bacterial drug resistance. We present a computational study of the structural and vibrational properties of clavulanic acid and tazobactam intermediates, two important β -lactamase inhibitors, bound to the singly mutated E166A SHV β -lactamase in aqueous solution by hybrid molecular mechanics/quantum mechanics (QM/MM) simulations at ambient conditions. We compare the Raman spectra obtained from the time autocorrelation function of polarizability tensor as obtained from a QM/MM protocol to those obtained from the instantaneous normal modes analysis performed on top of the QM/MM trajectory in order to establish the accuracy of these two computational methods and to review the previously made assignments. It is shown that the O=C–C=C–NH– trans-enamine moiety symmetric and asymmetric stretchings are strongly coupled with the N–H in-plane rocking and originate the band structure between 1600 cm^{-1} and 1640 cm^{-1} . Results indicate also that to properly describe Raman scattering properties of the trans-enamine intermediate, it is crucial to include both mechanical (beyond the second derivative of the potential energy at equilibrium) and electrical (beyond the first derivative of polarizability) anharmonicity. In addition, we show that the environment electrostatic field dynamically modulates the Raman activity, enhancing or inhibiting it.

1. Introduction

Raman microscopy and spectroscopy have been extensively employed to study relevant problems of biological interest, such as the β -lactamase mediated bacterial resistance to antibiotics (refs 1–5 and references therein), a major clinical problem worldwide. Inhibitors are usually employed to enhance the therapeutic efficacy of specific antibiotics, but several aspects of the involved reactions are still to be understood.

Raman microscopy together with X-ray crystallography, in particular, establishes a correlation between the activity and the degree of stability of the trans-enamine intermediates formed during the inhibition reaction by clavulanic acid,

tazobactam (Figure 1a), and sulbactam, the only three inhibitors clinically available so far.⁶ The majority of studies looked at the inhibitor intermediates formed with E166A SHV, a mutated β -lactamase, which prevents the deacylation of the enzyme allowing trapping of the intermediates for X-ray crystallographic investigations (Figure 1b). A very recent Raman crystallographic study of the reaction intermediates from wildtype SHV has also been reported.⁵ Difference Raman spectra were obtained subtracting from the recorded spectrum of the inhibitor bound to the protein that of the protein alone.^{1,4,5} The assignment of the strong band observed in the difference Raman spectra at 1592, 1599, and 1605 cm^{-1} for tazobactam, sulbactam, and clavulanic acid, respectively, to the trans-enamine stretching mode was crucial to establishing such a correlation. The assignment was made on the grounds of *ab initio* calculations performed on simplified gas phase models and confirmed experimentally by the observed shift of the same peak in D₂O due to isotopic

* Corresponding author e-mail: raugei@sissa.it.

[†] SISSA and INFN-DEMOCRITOS center via Beirut 2.

[‡] Louis Stokes Cleveland, Department of Veterans Affairs Medical Center.

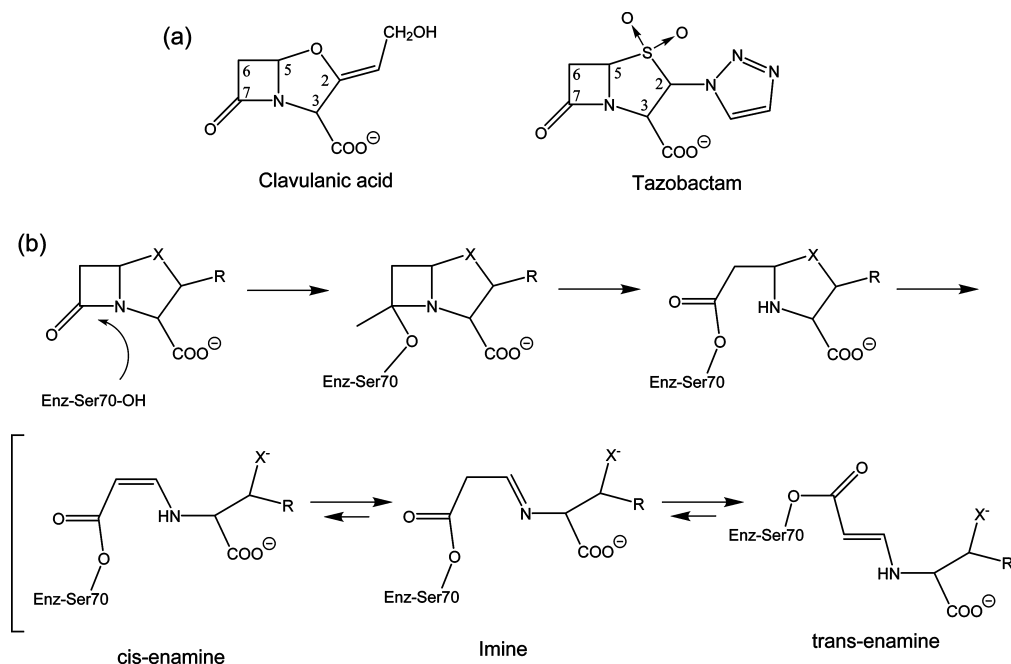


Figure 1. (a) Chemical structure of the two β -lactamase inhibitors considered in the present study (clavulanic acid and tazobactam). (b) Proposed reaction scheme for beta-lactam-based inhibitors with the deacylated deficient E166A SHV β -lactamases. Notice that, in the case of clavulanic acid, the reaction further proceeds with the intermediate decarboxylation.

substitution. The assignment of the observed isotopic shift, on the other hand, was again done following suggestions deriving from *ab initio* on the same model. Despite the number of investigations reported in the literature, the detailed understanding of the spectral features, especially in the region around 1600 cm^{-1} , has not been achieved yet (see for instance refs 1, 4, and 5). The purpose of this work is dual: (i) contribute to the assignment of the SHV β -lactamase/inhibitor Raman spectra and (ii) assess the accuracy achievable with the available quantum-chemistry methods for calculating vibrational spectra for complex systems of biological interest.

The calculation of vibrational spectra for solvated systems of biological interest is a very demanding task. Standard classical force fields are not sufficiently accurate to reproduce the experimental vibrational spectra. On the other hand, quantum mechanics (QM) simulations are still too computationally demanding. Even if QM methods could be employed for simulating the system, the calculation of the dielectric (polarizability) tensor for the whole system (or of a part of it) is problematic. For clavulanic acid in aqueous solution, we have recently shown⁷ that hybrid quantum mechanics/molecular mechanics (QM/MM) methods, in which the solute is treated quantum mechanically and the solvent with an empirical potential, are able to reproduce both the solute solvation and vibrational structures when compared to the more demanding full QM simulation, in which both the solute and the solvent are treated quantum mechanically. Moreover we have proposed a simplified and computationally efficient protocol to calculate the polarizability tensor for the whole system.⁷ In this work, we employ the same QM/MM simulations to calculate the Raman spectra of *h*- and *d*-clavulanic acid and *h*- and *d*-tazobactam trans-enamine intermediates bound to the E166A SHV β -lactamase, using the time autocorrelation functions (ACF) for-

malism. We then compare the QM/MM spectra to those obtained from the Instantaneous Normal Modes Analysis (INMA),^{8–10} probably the only other available method so far for calculating (routinely) vibrational spectra for such complex systems, in order better understand and compare the advantages and the limitations of techniques for the calculation of Raman spectra of complex systems.

The paper is organized as follows: in the next section (Section 2) the computational details of the methods employed in this work are explained. In Section 3, INMA and QM/MM results are discussed and compared to the available experimental data. Particular attention is dedicated to possible sources of error in the calculation of spectra, especially to statistical sampling errors. Finally, the main points of our study will be summarized in Section 4.

2. Computational Methods

2.1. Computational Models. The clavulanic acid and tazobactam intermediates bound to the deacylation-deficient E166A SHV β -lactamase (PDB structures 2A49 and 1RCJ, respectively) have been studied by long time scale classical molecular dynamics (MD) simulations based on the Amber¹¹ force field and TIP3P¹² water at atmospheric pressure and room temperature, followed by hybrid density functional theory (DFT)/molecular mechanics simulations.

The clavulanate and tazobactam-E166A SHV crystal structures include one and two nonstandard molecules, respectively: HEPES (4-(2-hydroxyethyl)-1-piperazin ethane-sulfonic acid) used as buffer and Cymal 6 (cyclohexyl-hexyl-beta-D-maltoside, MA4 hereafter) used as nonionic detergent in the crystallization procedure. HEPES is present only in the 2A49 structure and is found adjacent to the intermediate, being the sulfonic group near the ester group of the trans-enamine (Figure 2). The 1RCJ structure does not contain

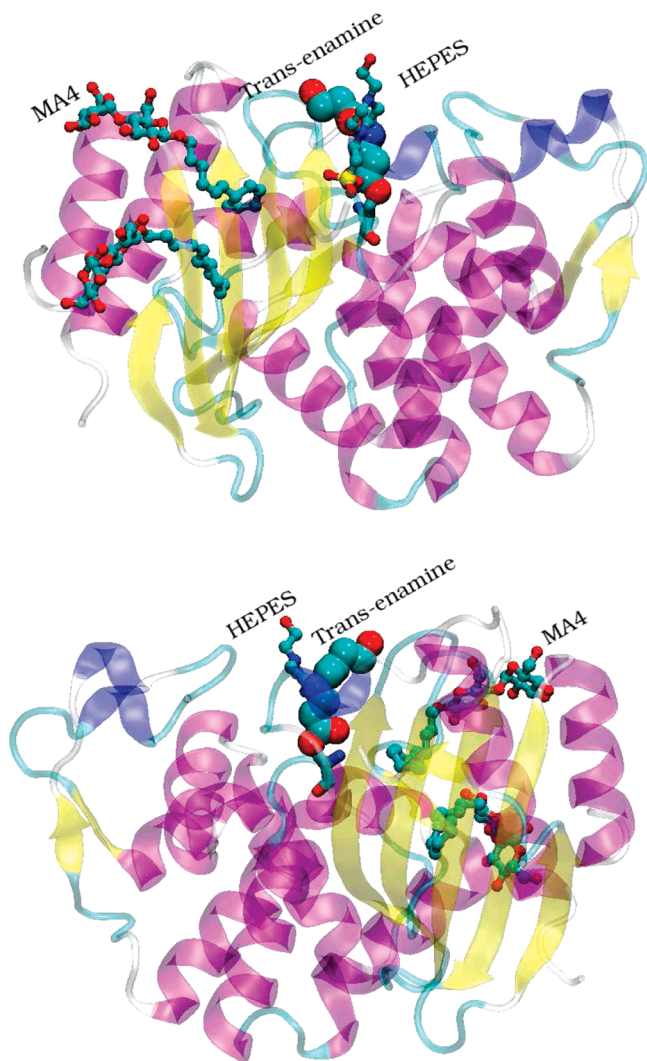


Figure 2. Two different views of the trans-enamine SHV β -lactamases intermediate crystallographic structure (PDB entry 2A49). The protein scaffold is shown in a cartoon representation, the trans-enamine moiety as spheres, the HEPES and the MA4 molecules in ball and sticks.

this molecule likely because of the bulky triazolyl moiety in tazobactam. In both crystal structures there are also two MA4 intercalated between helix 10 and helix 11 at more than 13 Å from the active site. Simulations were performed including these molecules. We remark that test simulations carried out without the MA4 molecules indicated that their insertion does not alter neither the protein structure or its flexibility. A further difference between the two crystal structures is the absence of the C3 carboxylate group in the intermediate from clavulanic acid (Figure 1).

The experimental Raman spectra were recorded during inhibitor soaking experiments for clavulanate and tazobactam¹ on a E166A SHV β -lactamase microcrystal. Simulations and the subsequent calculation of the Raman spectra were performed in an aqueous solution-like environment. This choice was dictated by the larger computational effort required to study the crystal phase as the crystal unit cell contains four protein molecules. On the other hand, although the catalytic pocket is at the protein surface, the trans-enamine intermediate is buried into the protein scaffold and

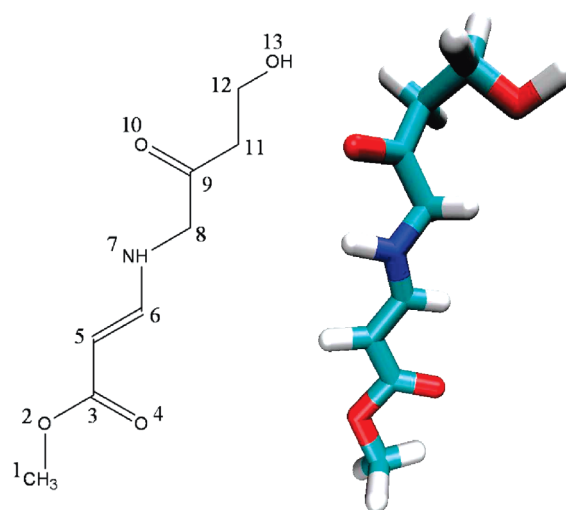


Figure 3. 2D and 3D structure of the quantum model used to describe the clavulanate trans-enamine intermediate together with the labeling of the heavy atoms employed in this work. The quantum region is terminated with the methyl group at position 1.

only the terminal part of the chain is exposed to water. In the crystal, the tail of the trans-enamine is separated from the nearest protein molecule by at least four water layers. Therefore, the major electrostatic influence to the static polarizability of the intermediate comes from the protein, the HEPES molecule (in the case of the intermediate from clavulanic acid), and the surrounding solvent. These facts make us confident that the spectrum calculated on a solution model of the protein adduct well represents the spectrum recorded on the microcrystal.

2.2. Classical Molecular Dynamics Simulations. The clavulanate and tazobactam intermediates as well as HEPES and MA4 were parametrized using the General Amber Force Field (GAFF),¹³ whereas the Restrained ElectroStatic Potential (RESP) charges were calculated at density functional theory level (B3LYP/6-31+G(d))¹⁴ using Antechamber¹⁵ and Gaussian.¹⁶ The proteins were immersed in an orthorhombic box containing 14082 and 13861 TIP3P¹² water molecules for clavulanate and tazobactam, respectively, plus sodium counterions to guarantee electroneutrality. The systems were first equilibrated for about 0.5 ns letting only water relax at room temperature. Then all the atoms were left free to move, and the temperature was gradually increased from 0 K to room temperature. The simulation was then continued at ambient conditions for 9 ns employing the Berendsen thermostat and barostat.¹⁷ The Amber package¹¹ was used to equilibrate the systems, whereas the Gromacs program was subsequently employed¹⁸ to run the production dynamics.

2.3. Hybrid QM/MM Molecular Dynamics Simulations. Starting from a snapshot of the classical MD simulations, a constant volume hybrid Quantum Mechanics–Car–Parrinello (CP)/Molecular Mechanics^{19–22} simulation was performed at room temperature partitioning the system into a quantum part, which includes the inhibitors and the Ser70 atoms up to the C_{β} (Figures 3 and 4), and a classical part, including water, buffer and detergent molecules, and the remaining part of the protein. The QM region was described within gradient-corrected DFT using the Becke²³

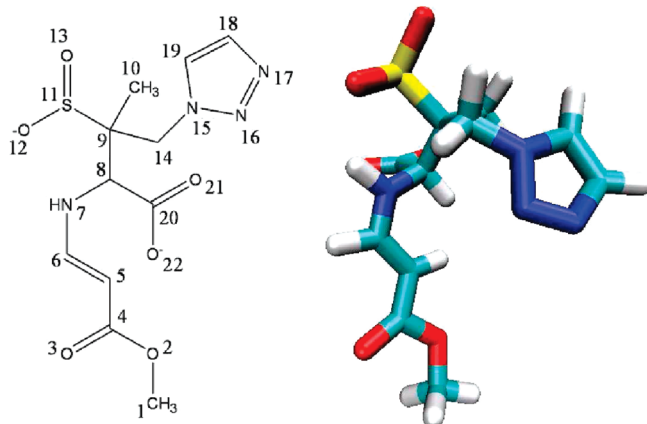


Figure 4. 2D and 3D structure of the quantum model used to describe the tazobactam trans-enamine intermediate together with the labeling of the heavy atoms employed in this work. The quantum region is terminated with the methyl group at position 1.

and Lee, Yang, and Parr²⁴ (BLYP) exchange and correlation energy functional. Norm-conserving Martins-Troullier pseudopotentials²⁵ were used to describe the interactions between the core and valence electrons for all atoms but for hydrogens, for which a von Barth-Car pseudopotential²⁶ was used to smooth out the short-range Coulombic nuclear potential. The Kohn–Sham orbitals of valence electrons were expanded in plane waves with kinetic energy up to 70 Ry. The electrostatic coupling between the QM and the MM region is calculated by using the full Hamiltonian scheme proposed by Rothlisberger and co-workers^{21,22} as implemented in the CP code CPMD 3.10.²⁷ Explicit electrostatic interactions between the quantum and the classical part were taken into account for residues and water molecules having at least one atom within 5.3 Å from any atom of the inhibitors. Electrostatic interactions between dynamically fitted D-RESP point charges²² located on the QM atoms and the force field RESP²⁸ charges on the MM atoms were considered for the remaining MM atoms within 10.6 Å from the QM part. Finally, electrostatic interactions between the MM atoms and a multipole expansion of the QM charge distribution were considered for MM atoms beyond 10.6 Å. The Car–Parrinello equations of motions were integrated using a fictitious electronic mass of 600 au and a time-step of 0.12 fs. The system was equilibrated for 1 ps with a Nosé–Hoover chain of thermostats,²⁹ and then the trajectory was sampled in the microcanonical ensemble for about 10 ps for the intermediate from tazobactam and about 24 ps for the intermediate from clavulanic acid taking advantage of its smaller size and therefore of its minor computational cost.

2.4. Raman Spectra. The vibrational Raman activity was calculated both via the standard quantum chemistry INMA method^{8–10} and the time ACF formalism with a QM/MM protocol previously established.⁷ In the latter, the time evolution of the polarizability tensor was calculated from the QM/MM trajectory extracting one configuration every 3.0 fs for clavulanate and 2.4 fs for tazobactam. These time intervals ensure the proper sampling of the mode with shortest vibrational period (that of N–H stretching vibration). The polarizability tensor for the chosen molecular configura-

tions was then evaluated using the BLYP density functional and the cc-pVDZ basis set,³⁰ including the effect of the surrounding electrostatic potential due to the solvent and protein.^{31,32} We included the charges of the whole protein and of a spherical shell of 25 Å of TIP3P water molecules around the inhibitors as obtained from the QM/MM trajectory. The same exclusion rules for the electrostatic interactions employed in the QM/MM simulation were used. It was previously noticed that, in the case of aqueous solutions, the polarizability tensor is not significantly affected when limiting the calculation to two shells of charges only.⁷ The calculation of the polarizability tensor for each snapshot required about 30 min for clavulanate, and 1 h and 25 min for tazobactam on one core of an AMD Opteron Dual Core (2.6 GHz) processor with the Gaussian03 package.¹⁶ To reduce the required computational time for tazobactam without significantly affecting the accuracy of the calculated polarizability, we employed the pseudopotential LanL2ZD for the sulfur atom.³³

As previously reported,⁷ the following expression for the Raman scattering cross section per unit of solid angle and unit of frequency, $I_{\perp}(\omega)$, observed perpendicularly to the incident light of frequency ω_0 can be written as^{34,35}

$$I_{\perp}(\omega) \propto (\omega_0 - \omega)^4 \int dt e^{-i\omega t} \langle \text{Tr} \beta(0) \cdot \beta(t) \rangle_{qm}^+ \quad (1)$$

$$= (\omega_0 - \omega)^4 C_{qm}(\omega)$$

with β being a traceless anisotropic part of our systems polarizability, α , calculated in a molecular fixed frame, $\beta = \alpha - \bar{\alpha} \mathbf{I}$, where $\bar{\alpha}$ is the average of the polarizability tensor trace, and \mathbf{I} is the unit tensor. After diagonalizing the polarizability tensor, the Raman scattering activity can then be obtained by calculating the Fourier transform of its symmetrized quantum mechanical ACF, $C_{qm}(\omega)$, which, following Bader and Berne,³⁶ is approximated in terms of the corresponding classical ACF, $C_{cl}(t)$, as

$$C_{qm}(\omega) = \frac{\hbar\omega}{2k_B T} \coth\left(\frac{\hbar\omega}{2k_B T}\right) C_{cl}(\omega) \quad (2)$$

where \hbar is the reduced Planck constant.

The vibrational Raman spectra for our systems were also obtained using the INMA protocol.^{8–10} INMA vibrational modes are calculated for an instantaneous relaxed configuration in the confinement potential of the environment. For this reason, they are often referred to as “quasi-harmonic”. Thirty-two frames equispaced in time were selected from the QM/MM trajectories of *h*-, *d*-tazobactam and clavulanate and from the classical trajectory for *h*- and *d*-clavulanate. For each frame, the QM subsystem was then relaxed keeping the classical part of the system fixed at the given geometry and gradually annealing the quantum ions up to a residual force on each QM atom of 10^{-5} a.u. The harmonic frequencies and the derivatives of the polarizability tensor for the optimized structures were then obtained using Gaussian,¹⁶ including the external charges of the remaining system, for the vibrational analysis. The calculation was performed by using the cc-pVTZ basis set. The BLYP exchange and correlation functional was employed. In the case of the

intermediate from clavulanic acid, BLYP results were verified against the more accurate hybrid B3LYP exchange and correlation functional.¹⁴ Harmonic Raman activities were obtained from the expression

$$\tilde{I}_{\perp}(\omega_i) = 45\alpha'(\omega_i)^2 + 7\beta'(\omega_i)^2 \quad (3)$$

where \tilde{I}_{\perp} refers to the total Raman scattering activity determined by the perpendicular component of the incident light electric field with respect to the scattering direction. The terms α' and β' appearing in \tilde{I}_{\perp} are defined as

$$\alpha' = \frac{1}{3}(\alpha'_{x,x} + \alpha'_{y,y} + \alpha'_{z,z}) \quad (4)$$

and

$$\beta'^2 = \frac{1}{2}[(\alpha'_{x,x} - \alpha'_{y,y})^2 + (\alpha'_{x,x} - \alpha'_{z,z})^2 + (\alpha'_{y,y} - \alpha'_{z,z})^2 + 6(\alpha'_{x,y}{}^2 + \alpha'_{x,z}{}^2 + \alpha'_{y,z}{}^2)] \quad (5)$$

where the prime indicates the derivative of the polarizability tensor elements with respect to the i -th normal mode.^{37,38} To calculate the relative Raman cross sections for the harmonic modes in the Stokes spectral region, the Raman scattering activity has been corrected with the factor

$$I_{\perp}(\omega_i) \propto \frac{(\omega_0 - \omega_i)^4}{\omega_i[1 - \exp(-\hbar\omega_i/k_B T)]} \tilde{I}_{\perp}(\omega_i) \quad (6)$$

where $\hbar = h/2\pi$ is the reduced Planck constant, k_B is the Boltzmann constant, and T is the absolute temperature. INMA spectra were simulated assuming a Lorentzian band shape with a bandwidth of 15 cm^{-1} . We remark that in ref 7 the formula for the intensity in the anti-Stokes region was erroneously reported. However in the calculations the proper expression for the Stokes region was employed.

Raman spectra for deuterated intermediates were also obtained via both QM/MM simulations and INMA. The calculation with QM/MM dynamics requires additional simulations where an H atom is replaced by a D isotope. In contrast, standard normal-mode analysis allows one to quickly evaluate isotopic effects by simply modifying the mass-weighted Hessian matrix since, in the Born–Oppenheimer approximation, the Hessian does not depend on the isotopic composition. We performed calculations for two different isotopic H/D substitutions. The first one concerns the enamine N–H hydrogen, which most easily undergoes H/D exchange with the solvent.^{1,5} In addition, also the possibility of a double deuterium exchange is taken into account. Indeed, it has been recently proposed⁵ that a second deuterium could be incorporated into the enamine skeleton at C6 (according to the numbering of Figure 1a or C5 according to the numbering used for the intermediates as in Figures 3 and 4). For computational reasons, the double substitution is taken into account only with INMA.

2.5. Sampling Issues. The limited sampling of the configurational space is source of error in the calculation of the Raman spectrum by using both ACF and INMA. To give an estimate of the uncertainty in both position and intensity of the bands, the Raman spectrum for the clavulanate

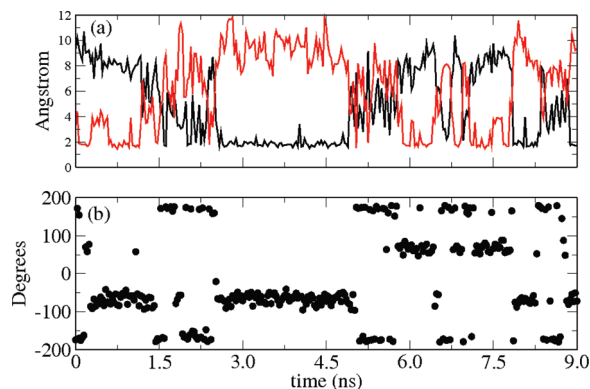


Figure 5. Correlation between the distance between the OH group of the clavulanate intermediate and the Glu240 (black line) and Asp104 (red line) carboxylate oxygens, and the dihedral angle C9-C11-C12-O13 of the clavulanate intermediate. Atom numbering is according to Figure 3.

intermediate was evaluated for the two halves of the QM/MM trajectory. The results are reported as Supporting Information (Figure SI-1). Given the same overall sampling time window, INMA spectra converge faster than those from ACF in that the intensity of only a few bands are affected significantly by the number of configurations taken into account. The same kind of uncertainty is obtained by taking half (16) or, for some modes, a quarter (8) of the INMA points along the entire QM/MM trajectory or taking points on shorter trajectory segments (up to 6 ps). The spectrum calculated from the ACF is more sensitive to the length of the trajectory. This is mainly due to the statistical error on the time ACF. The bands most affected are those at 1590, 1280, and 1160 cm^{-1} , for which variations as high as 25% are observed between the spectrum calculated over the entire simulation and the 12 ps trajectory segments. However, the general shape of the spectrum, in particular the relative intensity of the bands, does not change taking a shorter trajectory. Unacceptable differences among spectra appear when only 6 ps-long trajectory segments are employed for the calculation of the polarizability ACF. On the basis of these considerations, we expect that the intensities of QM/MM spectrum calculated for the tazobactam intermediate, which has been evaluated from a 10-ps long trajectory, might be affected by a larger error than 25%.

3. Results and Discussion

3.1. General Structural Properties. The intermediate from tazobactam is mostly rigid. No conformational changes were observed during the classical and QM/MM simulations. A different situation is observed for the intermediate from clavulanate, where the dihedral angle C9-C11-C12-O13 at the tail of the intermediate chain oscillates between three average values, -70 and $+70$ and $+180$ degrees. These values correspond to three different conformations in which the terminal OH group of the clavulanate intermediate forms a hydrogen bond either with the carboxylate of Asp104 (-70 degrees), or with the carboxylate of Glu240 ($+70$ degrees), or with water (180 degrees). In Figure 5 the correlation between the variation of the distances between the clavulanate OH and the carboxylate groups of Glu240 and Asp104

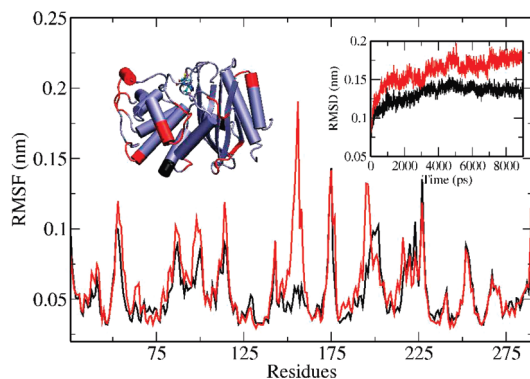


Figure 6. Root mean square fluctuations of the protein residues α carbons for E166A SHV tazobactam (black line) and E166A SHV clavulanate complexes (red line), together with the corresponding RMSF for the two proteins. In the inset regions with higher mobility are highlighted in red and black for clavulanate and tazobactam intermediate protein structure, respectively.

is reported, along with the corresponding value of the dihedral C9-C11-C12-O13. The dihedral angle corresponding to the more stable configuration seems to be that forming a hydrogen bond with Asp104. In fact, the carboxylate of Asp104 is constrained due to another hydrogen bond with the HEPES molecule, whereas that of Glu240 can freely rotate. This is also the conformer that features the smallest root-mean-square displacement of main chain heavy atoms from the crystallographic structure ($0.32 \pm 0.26 \text{ \AA}$ compared with $0.53 \pm 0.31 \text{ \AA}$ and $0.59 \pm 0.29 \text{ \AA}$ of the other two conformations). For this reason, as starting point of our QM/MM simulation of the clavulanate intermediate, we chose a configuration in which the OH of the clavulanate intermediate forms an hydrogen bond with the Asp104 residue. We remark that the larger backbone rigidity in the tazobactam intermediate is likely due to the presence of the carboxylate group and the bulky triazole moiety.

The higher mobility of the clavulanate intermediate slightly affects the flexibility of some regions of the protein. This is shown in Figure 6 in which the root-mean-square fluctuations (RMSF) of the C_{α} for the residues of the E166A SHV clavulanate intermediate (red curve) and tazobactam intermediate (black curve) complexes are reported, together with the corresponding rmsd taken along the 9 ns classical MD trajectories. The regions that present a higher RMSF for the C_{α} are residues located in loops and helices far from the active site (inset of Figure 6).

Finally, it is worth pointing out that the trans-enamine $O=C-C=C-NH-$ moiety is nearly planar in the intermediate from clavulanic acid, whereas it deviates appreciably from planarity in the intermediate from tazobactam. The dihedral angle around the $C=C$ bond calculated from the QM/MM simulations is $180^{\circ} \pm 9^{\circ}$ and $173^{\circ} \pm 8^{\circ}$ for clavulanate and tazobactam intermediates, respectively (in the X-ray structure the same angle amounts at 180° and 168°).

3.2. Raman Spectra. In this section the calculated Raman spectrum of the E166A SHV trans-enamine intermediate

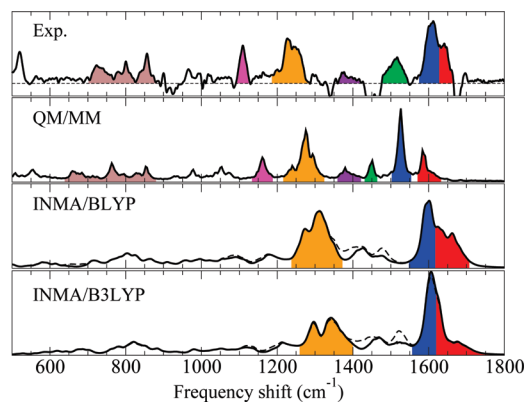


Figure 7. Experimental (ref 1), QM/MM, and INMA (BLYP and B3LYP, see text) Raman spectra for the *h*-clavulanate intermediate bound to the E166A SHV β -lactamase. The experimental data refer to the [protein + inhibitor] – [protein] difference spectrum (the dashed line indicates the zero intensity axis). Band colors indicate the correspondence between calculated and experimental bands based on the theoretical assignment. For INMA the spectrum calculated with (dashed line) and without (solid line) the contribution due to the methyl capping group of the QM subsystem are reported. Intensity in arbitrary units.

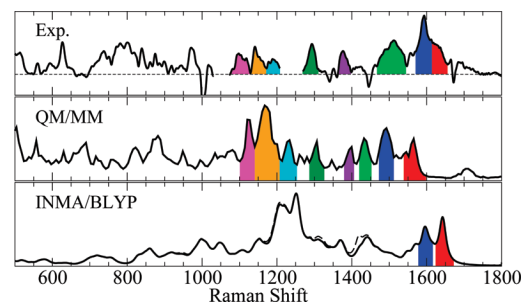


Figure 8. Experimental (ref 1), QM/MM, and INMA Raman spectra for the *h*-tazobactam intermediate bound to the E166A SHV β -lactamase. See also the caption to Figure 7.

from clavulanic acid and tazobactam will be discussed, and an interpretation of the main experimental features will be given.

3.2.1. General Spectral Features. The difference Raman spectrum for the two intermediates is reported in Figures 7 and 8. The spectral region of interest is the one below 1800 cm^{-1} . This region contains many peaks of varying intensities, many of which are not easily identifiable. We recall that positive bands in the [protein + inhibitor] – [protein] difference spectra reflect the [protein + inhibitor] state during or after the inhibition reaction, whereas negative bands are characteristic of the initial [protein] state.

The difference spectra are characterized by a high frequency band structure (around 1600 cm^{-1}) and by a broadband immediately lower in frequency (at 1515 cm^{-1} and 1512 cm^{-1} for clavulanate and tazobactam, respectively). The high frequency structure results from the overlap of two bands (at 1605 cm^{-1} and 1643 cm^{-1} for clavulanate and at 1595 cm^{-1} and 1630 cm^{-1} for tazobactam). In both cases, the band at around 1600 cm^{-1} has been taken as the signature of the enamine species. The band is slightly red-shifted upon NH deuterium exchange (1587 cm^{-1} for *d*-clavulanate and

1582 cm^{-1} for *d*-tazobactam). In addition, the band at 1510 cm^{-1} disappears upon deuterium exchange. The difference spectrum recorded for sulbactam (not considered in the present study) also reports these features.

The spectra calculated within the QM/MM ACF formalism feature the same number of positive bands as the experimental spectra with approximately the same relative intensity (Figures 7 and 8). This allows us to make a one-to-one correspondence between calculated and experimental bands as shown in the figures. As can be noticed immediately, the three high frequency bands are largely red-shifted with respect to the recorded spectra. In addition, the relative splitting of the bands is larger in the simulated QM/MM spectra. The position of these bands in the recorded trans-enamine clavulanate intermediate is at higher frequency than for tazobactam. This feature is reproduced by the calculated spectra even though the difference between clavulanate and tazobactam bands is more pronounced.

The Raman spectra simulated from the ACF-protocol are also able to reproduce the band structure between 1100 cm^{-1} and 1300 cm^{-1} although, in contrast to the high frequency modes, the calculated frequencies are shifted to the blue. In contrast, the spectra calculated from INMA on top of the QM/MM trajectory (Figures 7 and 8) only reproduce a few experimental features well, namely the high frequency bands (above 1580 cm^{-1}) and the large band structure around 1250 cm^{-1} . Other bands are very weak and not easily discernible. In particular the experimental band around 1510 cm^{-1} is either extremely weak or not present at all.

One of the most serious limitations of INMA is the harmonic approximation on which the method relies. On the other hand, the QM/MM spectrum calculated from the ACF fully takes into account both mechanical and electrical anharmonicity. The only limitation is imposed by the underlying potential energy surface and, in particular, by the exchange and correlation functional employed, i.e., the BLYP exchange and correlation functional.^{23,24} This problem clearly affects in the same measure the QM/MM and the INMA spectra. It is well known that the BLYP functional generally underestimates harmonic vibrational frequencies (see for example ref 39). On the other hand, this functional has been proved to correctly reproduce the anharmonic effects in organic molecules.³⁹ As can be noticed from Figures 7 and 8, QM/MM and INMA frequencies are quite different. In particular INMA seems to overestimate the frequency of modes in the 1580–1000 cm^{-1} region with respect to both QM/MM and experiment. In addition, as already mentioned, in the QM/MM Raman spectra calculated within the ACF formalism, vibrational frequencies in the interval 1100–1300 cm^{-1} are also blue-shifted (Figures 7 and 8). Overall, this is in contrast with general deficiencies of the BLYP functional just mentioned. It is deemed therefore necessary to assess the overall reliability of BLYP functional for the present system. To this end, the INMA spectrum for the intermediate from clavulanic acid was also calculated employing the B3LYP functional,¹⁴ which performs well in vibrational calculations.³⁹ In Figure 7 the INMA spectra calculated with the two exchange and correlation functionals is shown (see also Table 3 where the detailed information about some

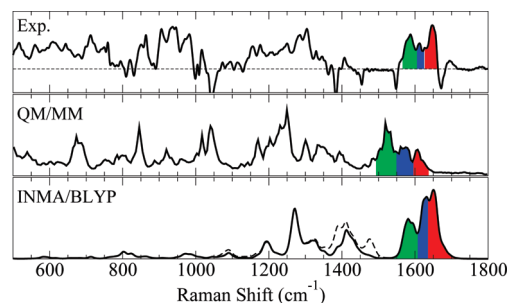


Figure 9. Experimental (ref 1), QM/MM, and INMA Raman spectra for the *d*-clavulanate intermediate bound to the E166A SHV β -lactamase. See also the caption to Figure 7.

vibrational mode is reported; the assignment of the Raman bands is discussed in the next section). As can be seen, in our specific case, when averaging over many configurations, the differences between the two functionals do not appear to be relevant. The observed frequency blue shift could be a result of the embedding empirical force field. In a previous work on clavulanic acid in aqueous solution,⁷ no appreciable frequency differences were reported between QM/MM and full QM frequency calculation. However, in the present case the effect on the vibrational properties of the (highly isotropic) embedding potential is difficult to predict *a priori*. The difference in the shift of the high frequency bands between QM/MM ACF and INMA spectra can be explained in terms of anharmonicity as discussed in Section 3.2.2.

Upon isotopic $\text{NH} \leftrightarrow \text{ND}$ exchange, the broad band at about 1515 cm^{-1} disappears. The calculated QM/MM spectra of the deuterated clavulanate trans-enamine also lacks the 1450 cm^{-1} peak (Figure 9b). The experimental clavulanate spectrum in D_2O reports also the separation of the two overlapping bands between 1570 cm^{-1} and 1670 cm^{-1} into three quite well resolved bands. This feature is also nicely reproduced by the calculated QM/MM spectrum for *d*-clavulanate (Figure 9b). We remark that the high frequency band is located near the amide I band of the unbound enzyme (negative band at 1670 cm^{-1}). In general, the enamine spectrum of clavulanate is more difficult to obtain because the enamine peaks are broader and overlap to a greater extent with the protein amide I band, which is subtracted away to generate the difference spectra. The residual amide I band intensity, which is present in both *d*- and *h*-clavulanate spectra, is likely due to small conformational changes in the protein on acylation of Ser70 residue or incomplete subtraction of the protein in the difference spectra. For these reasons, the positive band at 1648 cm^{-1} cannot be unambiguously assigned to the intermediate. The main difference between the calculated spectra for *h*- and *d*-tazobactam is the disappearance of the band at 1540 cm^{-1} (Figure 10b).

Further comparison with the experimental spectra of the deuterated trans-enamine intermediates is made problematic because of additional features, which appear in the difference spectra because of the difficulty of subtracting to zero intense peaks that are due to the buffer and the protein. In addition, to prepare the protein crystals in deuterated buffer, the crystals are removed from the mother liquor solution and undergo soaking for 24 h in advance of the Raman experi-

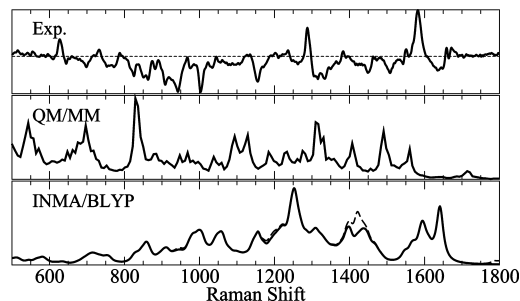


Figure 10. Experimental (ref 1), QM/MM, and INMA Raman spectra for the *d*-tazobactam intermediate bound to the E166A SHV β -lactamase. See also the caption to Figure 7.

ment to ensure H \leftrightarrow D exchange. This also influences the quality of the Raman difference spectra, introducing more noise.

We finally remark that, compared to experiments, the calculated QM/MM spectra feature narrower bands. This could be due to differences in temperature between experiments and numerical simulations. Indeed, the specimen temperature at the laser spot might be as high as 40 °C compared to room temperature at which simulations are carried out.

3.2.2. Vibrational Assignment. The assignment of vibrational spectra of large molecular assemblies represents a major problem. To this end, several methods have been proposed in the literature (see for instance refs 7, 8, 10, 40, and 41). In the present work a quantitative assignment has been performed using the vibrational density of states (VDOS), as obtained from the power spectrum of the velocity-velocity time autocorrelation functions, projected along the “quasi-harmonic” vibrational modes calculated from INMA. As discussed in a previous publication,⁷ many INMA eigenvectors are a good representation of a given vibrational mode at finite temperature. Indeed, INMA modes do not lose their character appreciably as a consequence of anharmonicity, giving rise to sharp single peak spectral features when used to project atomic velocities (Figure 11). Instead, some other modes are strongly mixed up by anharmonicity. This is particularly relevant for modes localized on very flexible parts of the trans-enamine intermediates. For this reason, using this projection procedure, it is difficult to assign the low frequency modes (<1000 cm⁻¹). However, this spectral region is not necessary to gain an understanding of the inhibitor/lactamase reaction mechanism. Overall, even though in the present case INMA does not seem to reproduce the overall Raman spectrum as well as the QM/MM simulation, the eigenmodes that are obtained represent a good basis set for band assignment. In Figure 11 some examples of projections for modes above 1100 cm⁻¹ are reported for the clavulanate intermediate.

The projection procedure allows us to assign several important bands of the Raman spectra. In Tables 1 and 2 the results of this analysis are summarized for clavulanate and tazobactam, respectively. Although many instant normal modes conserve their identity during the intermediate dynamics, their vibrational frequencies largely differ from the harmonic one. In particular the vibrational modes between 1580 and 1700 cm⁻¹ are downshifted systematically in the

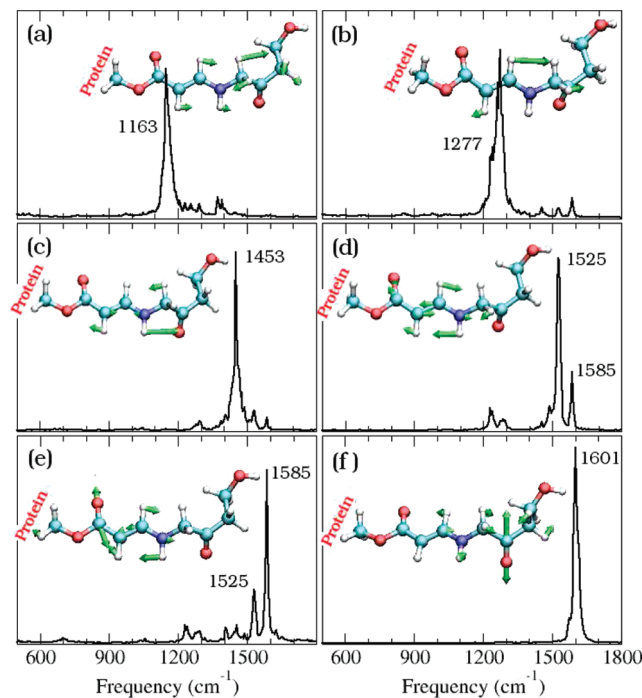


Figure 11. Vibrational density of states projected along a given Raman active (instant normal) mode obtained from the INMA. In each panel, the corresponding normal mode is shown with green arrows (only atomic components larger than a cutoff value $c = 0.1$ are reported). Intensity in arbitrary units. The calculation has been performed projecting velocities on atoms with eigenmode components greater than c . Each molecular dynamics frame has been roto-translated in a such a way to minimize the root-mean-square deviation from the configuration employed in the INMA analysis (only the atoms used for the projection have been considered in the minimization).

QM/MM simulation. This is not a consequence of the different computational framework used for the QM/MM simulation (pseudopotentials for the core functions and plane waves up to 70 Ry for the valence electrons) and the INMA (all-electron calculations with a triple- ζ cc-VTZ basis set). Indeed the harmonic frequencies of small model molecules (acetone, ethylene, and N-methyl enamine) calculated with the two methods barely differ (differences within 10 cm⁻¹). Therefore, discrepancies are likely due to anharmonicity, which is not taken into account in the INMA. For this reason the better accord between INMA and experiment in the frequency range from 1590 cm⁻¹ to 1650 cm⁻¹ should be considered fortuitous.

In the following, the vibrational assignment is discussed for the region from 1000 cm⁻¹ to 1800 cm⁻¹ and in particular for the region from 1500 cm⁻¹ to 1800 cm⁻¹, which has been proved critical for the understanding of the nature of the SHV reaction intermediate. The intermediate from clavulanate and tazobactam will be discussed in parallel as much as possible since they present common spectral features due to the enamine moiety.

Enamine Related Modes. The two experimental bands around 1600 cm⁻¹ and 1640 cm⁻¹ in the experimental spectra are unambiguously assigned to the C³=O⁴/C⁵=C⁶ symmetric (1640 cm⁻¹) and asymmetric stretching (1600 cm⁻¹) strongly

Table 1. Vibrational Analysis for *h*- and *d*-Clavulanate Bound to E166A- β Lactamase^a

<i>h</i> -clavulanate				<i>d</i> -clavulanate				assignment
QM/MM freq	INMA		exp. freq	QM/MM freq	INMA		exp. freq	
	freq	int			freq	int		
978	999(8)	2(2)		968	988(8)	3(2)		out-of-plane CH=CH-NH X-H atom wagging
1057*	1071(17)	5(3)		1021	1039(21)	5(2)		mostly C ⁸ -N ⁷ stretch
1075*	1100(28)	9(7)		1045	1086(11)	8(5)		C ⁵ -H in-plane rocking and C ⁸ H ₂ , C ¹¹ H ₂ twisting
1160*	1179(12)	9(5)		1171	1185(11)	8(5)	1105	mostly C ⁸ H ₂ twisting
1163*	1187(13)	6(4)		1179	1192(13)	6(5)	1105	coupled C ¹¹ H ₂ and C ¹² H ₂ twisting
1240*	1299(15)	18(15)		1234	1299(15)	15(14)	1202	mostly C ¹¹ H ₂ wagging and C ¹² H ₂ twisting
1277*	1310(13)	25(23)	1230	1250	1312(15)	6(5)	1230	in-plane, in-phase enamine CH=CH-NH-CH ₂ H-atoms rocking
1296*	1321(20)	18(17)	1295	1301	1321(20)	18(17)	1250	in-plane enamine CH=CH-N-CH ₂ H atoms rocking
1380*	1417(16)	17(5)		1387	1416(16)	18(10)		C ⁸ H ₂ scissoring
1393*	1422(13)	10(4)		1401	1424(8)	11(5)		C ¹¹ H ₂ scissoring
1410*	1465(11)	10(2)		1427	1465(11)	9(1)		C ¹² H ₂ scissoring
1450	1540(36)	3(1)		1447	1567(33)	5(4)		O ¹³ -H wagging
1453*	1511(15)	10(3)	1515	995	986(6)	4(3)		N ⁷ -H rocking (no activity for <i>d</i> -clavulanate)
1530*	1602(12)	76(22)	1605	1523	1586(15)	46(19)	1592	coupled symmetric C ³ =O ⁴ ; C ⁵ =C ⁶ stretching and C ⁶ -H; N ⁷ -H rocking
1588*	1652(18)	29(16)	1643	1570	1642(12)	40(22)	1613	coupled asymmetric C ³ =O ⁴ ; C ⁵ =C ⁶ stretching and C ⁶ -H; N ⁷ -H rocking
1605*	1682(17)	17(8)	-	1605	1682(16)	16(7)	1648	C ⁹ =O ¹⁰ stretching
1323	1089(11)	5(2)						CH ₃ capping group bending
1410	1385(12)	13(8)						CH ₃ capping group twisting
1415	1390(11)	13(10)						CH ₃ capping group umbrella
1730	1479(9)	16(3)						CH ₃ capping group bending

^a For instant normal mode analysis (BLYP/cc-pVTZ level), INMA, the average frequency and the Raman scattering activity (Int) are also given (standard deviations in parentheses). Frequencies in cm⁻¹ and Raman activities in Å⁴/a.m.u. QM/MM-Raman active modes are labeled with a dot (*). Atom numbering is according to Figure 3.

Table 2. Calculated and Experimental Raman Spectrum for *h*- and *d*-Tazobactam Bound to E166A- β Lactamase^a

<i>h</i> -tazobactam				<i>d</i> -tazobactam				assignment
QM/MM freq	INMA		exp. freq	QM/MM freq	INMA		exp. freq	
	freq	int			freq	int		
848	828(18)	2(3)		838	839(17)	1(3)		triazole C ¹⁸ -H, C ¹⁹ -H asymm. wagging
882*	931(37)	12(5)		882	934(19)	9(6)		S ¹¹ O ₂ symmetric stretching
925	993(6)	15(9)		930	992(12)	18(11)		S ¹¹ O ₂ asymmetric stretching
1125*	1206(16)	31(25)	1100	1128	1208(11)	34(20)		triazole N ¹⁵ -N ¹⁶ =N ¹⁷ asymm. stretching
1170*	1210(10)	66(23)	1143	1095	1163(26)	25(10)		in-plane, out-of-phase C ⁵ -H, N ⁷ -H rocking
1200*	1247(12)	73(37)		1200	1225(12)	20(12)		mostly C ⁸ -H bending
1275*	1234(15)	40(29)	1290	1269	1249(10)	79(26)		triazole breathing mode
1285*	1337(11)	7(3)		1280	1333(15)	11(7)		in-plane, in-phase C ⁶ -H and C ⁸ -H rocking
1292*	1282(9)	15(13)		1231	1279(13)	19(11)		C ²⁰ O ₂ carboxyl symmetric stretching
1304*	1374(9)	6(3)	(1290)	1316	1395(9)	41(16)		in-plane, enamine CH=CH-N-CH ₂ H atoms rocking
1348*	1368(10)	31(11)		1353	1373(11)	11(10)		C ¹⁰ H ₃ umbrella
1355*	1419(7)	14(9)		1348	1418(6)	14(8)		triazole N ¹⁵ =C ¹⁹ stretching
1398*	1442(7)	30(15)	1378		1442(7)	30(15)		C ¹⁰ H ₃ and C ¹⁴ H ₂ scissoring
1402	1409(4)	12(8)		1405	1456(9)	7(5)		C ¹⁰ H ₃ scissoring
1409	1457(9)	7(4)		1406	1457(9)	7(4)		triazole C ¹⁸ =C ¹⁹ stretching
1435*	1495(19)	21(12)	1512	(990)	1040(12)	6(3)		N ⁷ -H rocking (no activity for <i>d</i> -tazobactam)
1488	1594(16)	10(8)		1488	1595(22)	12(18)		C ²⁰ O ₂ carboxyl asymmetric stretching coupled to the C ⁸ -H wagging
1498*	1607(21)	49(44)	1595	1486	1581(18)	33(25)	1583	coupled symmetric C ³ =O ⁴ ; C ⁵ =C ⁶ stretching and C ⁶ -H; N ⁷ -H rocking
1562*	1629(21)	78(50)	1630	1548	1617(15)	69(59)		coupled asymmetric C ³ =O ⁴ ; C ⁵ =C ⁶ stretching and C ⁶ -H; N ⁷ -H rocking
1155	938(10)	4(3)						CH ₃ capping group bending
1260	1194(8)	13(10)						CH ₃ capping group twisting
1418	1349(7)	29(20)						CH ₃ capping group umbrella
1716*	1289(14)	11(6)						CH ₃ capping group bending

^a For instant normal mode analysis (BLYP/cc-pVTZ level), INMA, the average frequency and the Raman scattering activity (Int) are also given (standard deviations in parentheses). Frequencies in cm⁻¹ and Raman activity in Å⁴/a.m.u. QM/MM-Raman active modes are labeled with a dot (*). Atom numbering is according to Figure 4.

coupled with the C⁶-H/N⁷-H in-plane rocking (Figure 11d,e). These delocalized stretching/rocking modes are commonly referred to as O=C-C=C-NH- enamine backbone stretching.¹ In the QM/MM spectra, these two modes fall at 1530 cm⁻¹ and 1588 cm⁻¹ for clavulanate and

1488 cm⁻¹ and 1498 cm⁻¹ for tazobactam. These modes also have strong Raman activity according the INMA even though, as already remarked, they are higher in frequency and less separated (1601 cm⁻¹ and 1647 cm⁻¹ for clavulanate and 1607 cm⁻¹ and 1627 cm⁻¹ for tazobactam).

Table 3. Instant Normal Mode Analysis of the Effect of Isotopic Substitution on Some Stretching Vibrational Modes of SHV Reaction Intermediates from Clavulanic Acid and Tazobactam^a

asymm. (CO ₂)		asymm. (O=C-C=C-NH-)		symm. (O=C-C=C-NH-)		C=O	
freq	int	freq	int	freq	int	freq	int
Clavulanate BLYP							
0d		1602(12)	76(22)	1652(18)	29(16)	1682(17)	17(8)
1d		1586(15)	46(19)	1642(12)	40(22)	1682(16)	16(7)
2d		1581(16)	36(15)	1636(13)	61(20)	1682(16)	15(7)
Clavulanate B3LYP							
0d		1610(11)	109(27)	1675(14)	19(11)	1697(16)	9(1)
1d		1593(13)	87(23)	1655(12)	34(18)	1697(16)	10(2)
2d		1588(13)	77(24)	1647(13)	45(21)	1697(16)	10(2)
Tazobactam BLYP							
0d	1594(16)	10(8)	1607(21)	49(44)	1629(21)	77(51)	
1d	1595(22)	12(18)	1581(18)	63(48)	1617(15)	69(53)	
2d	1593(14)	19(21)	1576(16)	53(48)	1616(17)	80(54)	

^a Average frequency and the Raman scattering activity (Int) at BLYP/cc-pVTZ and B3LYP/cc-pVTZ level for normal (0d), singly deuterated at N7 (1d), and doubly deuterated at N7 and C5 (2d) intermediates are reported (standard deviations in parentheses). Frequencies in cm⁻¹ and Raman activity in Å⁴/a.m.u. Atom numbering is according to Figures 3 and 4.

In the QM/MM Raman spectrum of (singly deuterated) *d*-clavulanate, these two bands resolve into three overlapping peaks. Of these, the two peaks at lower frequency are assigned to the delocalized trans-enamine stretching and rocking mode. The highest frequency band instead is due to the pure enamine C⁹=O¹⁰ vibration (Figure 11f). This mode does not undergo any frequency shift upon isotopic exchange, and it might be responsible for the high frequency shoulder of the 1588 cm⁻¹ band in the *h*-clavulanate spectrum. We point out that the C⁹=O¹⁰ stretching presents an appreciable Raman activity also in the INMA (Table 1). The experimental spectrum of *d*-clavulanate also displays three bands. The one at high frequency (around 1650 cm⁻¹) is however doubtful as it is located near a possible protein amide I band subtraction artifact. In addition, its position is quite anomalous for a ketonic group being about 60 cm⁻¹ lower than the typical frequency of a saturated ketone C=O stretch. Nevertheless, the C=O mode has generally an appreciable Raman activity, and a signal due to it should be expected in the spectrum. However, no bands are present in the experimental spectrum of both *h*- and *d*-clavulanate in the 1700–1800 cm⁻¹ interval. For this reason, the possibility of a large redshift for the C=O stretching frequency was investigated. Several reasons can be envisaged for the large redshift of the CO band. It has been reported in literature that hydrogen bonding and the environment electric field have a bathochromic effect on the C=O stretch.^{42–44} In the present case, the C⁹=O¹⁰ group is found to be hydrogen-bonded to 1.1 ± 0.7 water molecules and to the Asn¹⁷⁰ side chain. In a first attempt to elucidate the role of these hydrogen bonds and of the electrostatic field of the protein and the aqueous medium generally, INMA was carried out in the absence of the classical point charges. The resulting C=O frequency was blue-shifted by about 35 cm⁻¹. To further address this point, the electronic structure of the carbonyl group was characterized in terms of localized Boys orbitals⁴⁵ (BO) corresponding to covalent bonds and lone pairs. The center of charge of the BOs (BOC) provides a useful tool to quantify the polarization of covalent bonds and to analyze differences in the electronic structure of the same functional group in different environments.⁴⁶ The electronic structure

of the intermediate in the enzyme pocket for 20 equispaced in time configurations was compared with that calculated, for the same configurations, in the absence of the environment electric field. In the first case, the BOC of the C=O double bond is located at 0.676 ± 0.004 Å from the carbon atom, whereas in the latter this distance is 0.657 ± 0.004 Å, very close to that in acetone in vacuo (0.656 Å). This sizable polarization of the carbonyl charge density is indicative of a larger weight of the carbonyl single-bond C⁺–O⁻ Lewis resonance structure to C=O bond and implies a reduction of the stretching vibrational frequency. This finding supports the idea that a shift of the intermediate C=O backbone frequency to the region of the amide I band is plausible, and it strongly supports the proposal that the experimental feature at 1650 cm⁻¹ is not an artifact but due to the C=O vibration.

We also point out that the symmetric and asymmetric enamine stretch modes can also account for the bands recently observed at 1595 cm⁻¹ and 1658 cm⁻¹ in the spectrum of the tazobactam intermediate from the wildtype SHV enzyme.⁵ On the basis of a vibrational analysis on minimal models of possible reaction intermediates in the gas phase,⁵ the latter was assigned to the N=H⁺ of an imine species coexisting with the trans-enamine intermediate (Figure 1). Our calculations suggest that the trans-enamine species also can explain this band.

Enamine Related H Rocking Bands. The experimental band around 1515 cm⁻¹ can be unambiguously attributed to the enamine N⁷–H in-plane rocking (Figure 11c). Indeed, this is the only vibrational mode below the enamine asymmetric stretching band common to both clavulanate and tazobactam intermediates. In the QM/MM spectra this mode is located at 1453 cm⁻¹ for clavulanate and 1435 cm⁻¹ for tazobactam and possesses a very large Raman scattering activity. However, the bandwidth in the QM/MM spectrum is considerably narrower than the experimental one. The INMA predicts little Raman scattering activity for this mode (in particular for the intermediate from clavulanate), but its frequency is more consistent with the experiment (1511 cm⁻¹ and 1495 cm⁻¹ for clavulanate and tazobactam, respectively). In the singly deuterated trans-enamine the N–H rocking

vibration remarkably downshifts at around 1000 cm^{-1} and does not show any appreciable Raman activity in either the QM/MM or the INMA spectra.

The enamine N–H rocking motion also gives rise to other bands in the Raman spectrum. In particular in the clavulanate intermediate the very intense band at 1277 cm^{-1} is mostly due to the in-plane, in-phase rocking motion of the H atoms of the $\text{C}^5\text{H}=\text{C}^6\text{H}-\text{N}^7\text{H}-\text{C}^8\text{H}_2$ fragment (Figure 11b). This band nicely correlates with the very intense peak at 1230 cm^{-1} in the recorded spectrum. The mode slightly redshifts upon deuterium substitution.

A similar vibrational mode has not been found in the tazobactam intermediate, likely because of the perturbation introduced by the bulky groups at C^8 . Rather a coupled $\text{C}^5-\text{H}/\text{N}^7-\text{H}$ mode gives rise to the very intense band at 1170 cm^{-1} , which correlates to the experimental band at 1100 cm^{-1} . Because of the large N–H contribution to this mode, the band downshifts of about 75 cm^{-1} in the deuterated QM/MM spectrum.

Finally, we remark that the experimental spectra of both inhibitors in aqueous solution and enzyme reaction intermediates present a band around 1290 cm^{-1} .¹ Tazobactam and the tazobactam intermediate feature a sharper and more intense band. In a previous study,⁷ the (broad) 1290 cm^{-1} band in the QM/MM spectrum of clavulanic acid in water was assigned (mostly) to the lactam ring C–H rocking mode (C^4-H according to Figure 1a numbering). The present theoretical analysis of the clavulanate and tazobactam reaction intermediates suggests that the C^5-H and C^6-H (which corresponds to the lactam ring C–H of the unreacted inhibitor) in-plane rocking mode has also an appreciable Raman scattering section and gives rise to a band around 1300 cm^{-1} . The unique assignment of the experimental feature around 1290 cm^{-1} band to the lactam or enamine C–H moiety is made problematic by the presence in this region of a triazolyl moiety mode.⁵ This point is further addressed when discussing tazobactam specific vibrational modes.

Clavulanate Specific Bands. The intermediate from clavulanic acid has three aliphatic CH_2 groups, which give rise to Raman activity in the QM/MM spectrum. The CH_2 scissoring modes originate the weak band at 1390 cm^{-1} , which correlates with the equally weak structure at 1375 cm^{-1} in the experimental difference spectrum. INMA also predicts Raman activity for the CH_2 scissoring modes.

In contrast CH_2 twisting modes show high activity in the QM/MM spectrum and are responsible for the intense band at around 1160 cm^{-1} , which we assign to the intense peak at 1105 cm^{-1} in the experimental spectrum. In addition, the C^{12}H_2 twisting coupled with the C^{11}H_2 scissoring gives rise to the peak at 1240 cm^{-1} to the left of the intense 1277 cm^{-1} band.

The modes so far discussed complete the band assignments of the QM/MM Raman spectrum of the enamine intermediate from clavulanic acid and provide a satisfactory interpretation of the experimental Raman spectrum between 1000 cm^{-1} and 1700 cm^{-1} .

Tazobactam Specific Bands. The intermediate from tazobactam has a more complex structure than that of clavulanate.

The most relevant differences are the presence of the triazole ring, the carboxylate group, and the sulfoxo group. These groups make the tazobactam Raman spectrum very rich.

Among all of the vibrational modes localized on the triazole ring only three present an appreciable Raman activity both in QM/MM and INMA above 1000 cm^{-1} . The first is the $\text{N}^{15}-\text{N}^{16}=\text{N}^{17}$ asymmetric stretching, which produces the band at 1125 cm^{-1} . This mode correlates well with the experimental band at 1100 cm^{-1} , even though its intensity is higher than the experimental one. The second Raman active mode is a triazolyl breathing-like mode (1275 cm^{-1} and 1234 cm^{-1} according the QM/MM simulation and the INMA, respectively). Its calculated QM/MM Raman scattering section is however little while quite large at the INMA level. The third active triazole mode is the $\text{N}^{15}=\text{C}^{19}$ stretching, which gives rise to the band at 1353 cm^{-1} . However its Raman activity is doubtful as it overlaps with the C^{10}H_3 umbrella motion at 1355 cm^{-1} .

In a recent investigation of the trans-enamine intermediate from wildtype SHV β -lactamase, Kalp et al.⁵ proposed that the 1290 cm^{-1} peak is due to the breathing mode of the triazolyl moiety. The present study on the E166A mutant is consistent with this assignment. However, as already discussed, C–H distortions also have a high Raman activity around 1300 cm^{-1} . Therefore, we suggest that both C–H rocking and triazole localized vibrations might both originate the observed band at 1290 cm^{-1} . This proposal explains why the 1290 cm^{-1} band in the intermediate from tazobactam has an intensity higher than that from clavulanic acid. We remark that on the basis of INMA Raman activities only the triazole localized mode could give rise to the 1290 cm^{-1} even though its frequency is considerably smaller than the experimental one.

The carboxylate symmetric stretch gives rise to a band at 1292 cm^{-1} with a weak intensity in the QM/MM spectrum (1326 cm^{-1} in the INMA with low intensity). The asymmetric CO_2 stretch falls at 1488 cm^{-1} , and it is either inactive or its intensity is lower than the background noise. However, in the INMA this mode shows a quite high activity at 1594 cm^{-1} . Kalp et al.,⁵ on the basis of gas-phase models, assigned the observed band at 1630 cm^{-1} in the spectrum of tazobactam intermediate from the wildtype SHV enzyme to the asymmetric CO_2 stretch. In their model, which lacks the triazolyl moiety and the sulfoxo group, the asymmetric CO_2 is strongly coupled to the enamine $\text{O}=\text{C}-\text{C}=\text{C}-\text{NH}-$ stretch, and its frequency down shifts upon NH/ND exchange according to observed behavior of the experimental band. However, according to our analysis, no such coupling between CO_2 symmetric stretch and enamine stretch is found, and, consequently, no appreciable isotopic effect is observed for the CO_2 stretch both in QM/MM and INMA spectra. The small structural differences between the wildtype enzyme and the E166A mutant here studied might suppress any coupling between the CO_2 symmetric stretch and enamine stretch observed for the gas-phase model. Our results suggest that the 1630 cm^{-1} in the wildtype SHV intermediate is due to some other species. Indeed, in light of the discussion about the enamine related modes above, a plausible candidate for the 1630 cm^{-1} is the imine intermediate (Figure 1).

The sulfone group stretching modes are strongly coupled with other stretching and bending modes all over the intermediates. The symmetric stretch gives rise to a moderately intense band at 882 cm^{-1} in the QM/MM, whereas the asymmetric mode (925 cm^{-1}) does not have any detectable Raman activity. In contrast, INMA predicts both modes to have an appreciable activity.

Finally, a vibrational mode due to the C^{10}H_3 and C^{14}H_2 scissoring unambiguously originates the band at 1398 cm^{-1} in the QM/MM. INMA predicts also a high activity for this mode. This completes the assignment of the Raman vibrational spectrum for the intermediate from tazobactam.

Comments on Isotopic H/D Exchange. In a recent report, it has been proposed that two deuterium atoms are incorporated into the enamine skeleton.⁵ Indeed, on the basis of a vibrational (harmonic) analysis of a minimal model of the trans-enamine intermediate in the gas-phase,⁵ the single enamine NH/ND exchange cannot explain the observed down shifts of the enamine $\text{O}=\text{C}-\text{C}=\text{C}-\text{NH}-$ (symmetric) stretching at 1595 cm^{-1} in the tazobactam intermediate from wildtype SHV. These calculations show that only when deuterium is included at C5 of the single intermediate skeleton (C6 of the lactam ring, Figure 1a) the enamine stretching shifts down in accord to experiments. In contrast, our vibrational analysis, carried out on the E166A SHV taking into account the entire intermediate and the coupling with the environment (protein and aqueous medium), indicates that the single NH/ND is sufficient to explain the observed isotopic shift. Indeed, the INMA results reported in Table 3 suggest that the C5 substitution does not induce any further appreciable frequency shift.

The discrepancies between the present calculations and those reported in ref 5 might be attributed to a reduced C^5-H rocking character and an increased N-H rocking character of the coupled $\text{O}=\text{C}-\text{C}=\text{C}-\text{NH}-$ trans-enamine symmetric stretching/H-rocking mode when passing from the minimal model to the full atomistic description of the intermediate at finite temperature (the ratio between the C^5-H and N-H characters goes from 14.0 to 0.2).

3.3. Effect of the Environmental Electric Field. The electrostatic field of the protein frame and the surrounding solution to which the trans-enamine is partially exposed obviously influence both frequency and Raman activity of the vibrational modes. We already mentioned how the environmental electrostatic field polarizes the $\text{C}^9=\text{O}^{10}$ carbonyl group of the intermediate from clavulanic acid, causing a redshift toward the amide I region. In this section we analyze in detail the role of the electrostatic embedding on the Raman spectrum. For computational reasons, we have focused our attention only on the intermediate from clavulanic acid.

In Figure 12, the spectrum calculated without the environmental electrostatic field is compared to the full spectrum. As expected, the electrostatic polarization of the intermediate due to the surrounding medium influences the intensity of several bands in both QM/MM and INMA spectra.

We will discuss the QM/MM spectrum first. As can be seen from Figure 12, two major bands are activated by the environment field. The first one is the band at 1163 cm^{-1} in

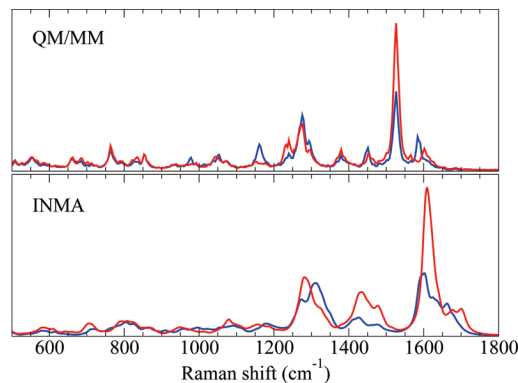


Figure 12. QM/MM and INMA Raman spectra calculated for the clavulanate intermediate with (blue line) and without (red line) the electrostatic fields of the environment. Intensity in arbitrary units.

the QM/MM spectrum, which corresponds to a coupled C^{11}H_2 and C^{12}H_2 twisting. The second and more important band is that at 1588 cm^{-1} , which is due to the asymmetric $\text{C}^3=\text{O}^4/\text{C}^5=\text{C}^6$ stretching coupled to the C^6-H ; N^7-H rocking. Conversely, the Raman intensity of other bands is reduced by the external electrostatic field. Among them, it is of interest to mention the band at 1530 cm^{-1} , which originates from the symmetric $\text{C}^3=\text{O}^4/\text{C}^5=\text{C}^6$ stretching coupled with the C^6-H ; N^7-H rocking and the band at 1600 cm^{-1} due to the $\text{C}^9=\text{O}^{10}$ stretching.

In the INMA spectrum no bands appear to be activated by the environment electrostatic field, which instead considerably reduces the intensity of (i) the CH_2 scissoring bands between 1410 and 1470 cm^{-1} , of (ii) the N-H rocking band at 1511 cm^{-1} , and, as for the QM/MM spectrum, of (iii) the symmetric $\text{C}^3=\text{O}^4/\text{C}^5=\text{C}^6$ stretching band at 1601 cm^{-1} .

Since the time evolution of the polarizability of the intermediate is modulated by the underlying dynamics of atoms, the position of bands in the QM/MM spectrum calculated from the ACF does not shift upon removal of the electrostatic field. This is of course not the case of the INMA spectrum. The larger frequency shift is observed for the three Raman active modes between 1600 cm^{-1} and 1650 cm^{-1} .

The comparison between QM/MM and INMA spectra calculated with and without the embedding electrostatic potential allows us to draw important conclusions on the role of environment modulation of the Raman activity. It is evident from Figure 12 that the mode at 1163 cm^{-1} (CH_2 twisting) is Raman active only in the spectrum calculated from the ACF and only when the (dynamical) modulation by the environment field is taken into account. This modulation is also responsible for the high Raman scattering intensity of the asymmetric $\text{C}^3=\text{O}^4/\text{C}^5=\text{C}^6$ stretching at 1588 cm^{-1} and of the small intensity of the $\text{C}^9=\text{O}^{10}$ stretching.

We finally remark that the INMA Raman activity for some modes show large fluctuations (i.e., high standard deviation from the average, see Tables 1 and 2). This is a consequence of the large change in the local electric field due to the surrounding environment. These changes are very rapid and depend mostly on the nearest coordination shells of the reaction intermediate. The large amplitude field oscillations have a characteristic time of a few picoseconds and average out on a multipicosecond time scale.

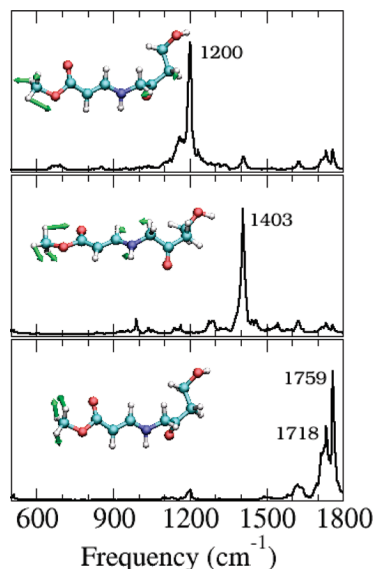


Figure 13. Vibrational density of states projected along three (instant normal) modes localized on the QM capping methyl group. Intensity in arbitrary units.

3.4. Influence of the QM/MM Interface: Effect of the Capping Atoms. The QM models have been built by cutting the inhibitor/protein adducts to the C_{β} of Ser70 and saturating its valence with an hydrogen atom. The resulting capping methyl group is linked to the MM part via standard bonded force field parameters (stretching, bending, and torsional energy terms). Of course this perturbs the intermediate electronic structure structure and its vibrational properties. In particular, vibrational modes involving the capping methyl group are artificially introduced. These modes might hamper the calculation of the Raman spectrum as they could be Raman active, and they could couple with other modes therefore altering their vibrational frequency. The latter issue cannot be easily addressed, but, in principle, it should be expected that the larger the coupling, the greater the perturbation.

Many INMA vibrational modes have an appreciable component on the capping group. Of these modes, the majority are localized on the methyl group and present, according to INMA, Raman activity in the region between 1300 cm^{-1} and 1500 cm^{-1} (see Figures 7–10 and Tables 1 and 2). In the INMA spectra discussed in the previous section, modes giving rise from this spurious Raman activity are neglected from the analysis. This is not done as easily in the QM/MM spectrum as the time evolution of the polarizability of the reaction intermediate is intrinsically modulated also by the capping group vibrations. However, no Raman activity has been found in the QM/MM spectra for modes localized on the capping group. In fact, as can be seen from Figure 13, in the intermediate from clavulanic acid, vibrational modes of the capping group present sharp spectral signatures when projected onto atomic velocities, and their frequencies do not coincide with any band in the Raman spectrum (Table 1). The same holds for the tazobactam intermediate (Table 2).

In addition, the frequency of these modes is largely blue-shifted with respect to INMA. This is due to the reduced

mobility of the capping hydrogen because of both nonbonded and bonded interactions via the force field. Indeed, these interactions make the confinement potential in which the capping hydrogen is located harder in the QM/MM simulation than that in the INMA. This also reduces the time modulation of polarizability by the motion of the capping hydrogen and therefore the influence on the Raman activity. Only the intermediate from tazobactam has one mode at 1716 cm^{-1} associated with the capping group with appreciable Raman activity (Figures 8 and 10).

4. Conclusions

The vibrational Raman spectra of the trans-enamine intermediate of the E166A SHV β -lactamase inhibition reaction by clavulanic acid and tazobactam, two major clinically available inhibitors, have been calculated from state-of-the-art atomistic calculations. Two methods have been compared: (i) quantum-chemistry harmonic vibrational analysis with temperature averaging included via instant normal-mode analysis (INMA)^{8–10} and (ii) power spectrum of the time autocorrelation function of the polarizability tensor components as obtained from QM/MM simulations. In both cases, the quantum mechanical problem was solved within the same DFT-theory framework. Differences are therefore expected to arise mainly from mechanical (frequencies beyond the second order derivatives of the potential energy) and electrical (band intensity beyond the first derivative of polarizability) anharmonic effects. Spectra calculated within the ACF formalism seem to better correlate with the experimentally recorded spectra as they feature the same number of bands with approximately the same relative intensity. In contrast, INMA seems to reproduce only a few experimental features. In particular, a band attributed to the enamine N–H in-plane rocking has very little Raman activity according to INMA. Other vibrational modes, which are active according to the spectra calculated from ACF, are missing in INMA. Instead, the spectra obtained from the QM/MM simulations allowed us to unambiguously assign the experimental spectra in the region between 1000 cm^{-1} and 1800 cm^{-1} . According to these simulations and consistent also with INMA calculations, the trans-enamine moiety gives rise to a $C^3=O^4/C^5=C^6$ symmetric and asymmetric stretching strongly coupled with the C^6-H/N^7-H in-plane rocking, which are Raman active and provide a rationale for the bands around 1600 cm^{-1} and 1640 cm^{-1} . Furthermore, a large redshift of the backbone carbonyl stretching frequency is predicted for the intermediate from clavulanic acid. Indeed, because of the polarization induced by the environmental electrostatic field, the $C=O$ frequency may nearly overlap with the amide I band and the asymmetric stretching mode delocalized on the trans-enamine moiety. Based on these findings, it is proposed that the $C=O$ stretch gives rise to a well resolved peak in the experimental spectrum of *d*-clavulanate intermediate.

It has been found that anharmonicity plays a major role in determining both frequencies and intensities, and it is the main reason for the discrepancies between INMA and QM/MM spectra. The importance of anharmonic effects in Raman (and also in infrared) spectra have been recently evidenced for alanine and proline zwitterions in solution (modeled as

a polarizable continuum).⁴⁷ The effects observed in the present study are however more pronounced than those reported for these amino acids. In addition, the comparison between the present QM/MM and INMA spectra calculated with and without the environment embedding potential clearly shows that the surrounding dynamically modulates the Raman activity, enhancing or inhibiting it. For instance, a vibrational mode due to the CH₂ twistings in the intermediate from clavulanic acid is Raman active only in the spectrum calculated from the ACF and only when the modulation by the environment field is accounted.

A major source of error in the calculation of the spectrum from QM/MM simulations comes from the limited phase space sampling, which implies (large) inaccuracies in the ACF. Nevertheless, hybrid QM/MM simulations combined with efficient protocols for the calculation of the polarizability of the target substrate seems to be a fast way to obtain a reasonably accurate description of the vibrational Raman effect. For example, the calculation of the Raman spectrum of the intermediate from clavulanic acid, from the preliminary classical force field simulations to the calculation of polarizability, took about 45 days on 16 cores of an AMD Opteron dual-processor dual-core cluster. In conclusion, our study demonstrates that a reliable and (computationally) efficient calculation of vibrational spectra of large biomolecular complexes can be (almost) routinely employed to help assign the experimental collected data. Probably more important, computational vibrational spectroscopy might become a precious tool for refinement of theoretical models when no other structural information is available.

We conclude remarking that, although the trans-enamine species is the most abundant reaction intermediate and the present study succeeds in interpreting most of the recorded spectral features, further calculations on other possible intermediates (imine and cis-enamine) are required for a more exhaustive understanding of the experimental spectra⁵ and, therefore, the inhibition mechanism.

Acknowledgment. The authors acknowledge computational support from INFM and CINECA. M.S.H. would like to acknowledge the Department of Veterans Affairs in Cleveland for the support of a Career Development Award.

Supporting Information Available: Estimate of the uncertainty in both position and intensity of bands in the calculated Raman spectrum for the clavulanate intermediate. This material is available free of charge via the Internet at <http://pubs.acs.org>.

References

- (1) Helfand, M. S.; Totir, M. A.; Carey, M. P.; A, M, H.; Bonomo, R. A.; Carey, P. R. *Biochemistry* **2003**, *42*, 13386–13392.
- (2) Helfand, M. S.; Taracila, M. A.; Totir, M. A.; Bonomo, R. A.; Buynak, J. D.; van den Akker, F.; Carey, P. R. *Biochemistry* **2007**, *46*, 8689–8699.
- (3) Carey, P. R. *Annu. Rev. Phys. Chem.* **2006**, *57*, 527–554.
- (4) Padayatti, P. S.; Helfand, M. S.; Totir, M. A.; Carey, M. P.; Carey, P. R.; Bonomo, R. A.; van den Akker, F. *Biochemistry* **2004**, *43*, 843–848.
- (5) Kalp, M.; Totir, M. A.; Buynak, J. D.; Carey, P. R. *J. Am. Chem. Soc.* **2009**, *131*, 2338–2347.
- (6) Bonomo, R. A.; Rice, L. B. *J. Gerontol A-Biol* **1999**, *4*, B260–B267.
- (7) Miani, A.; Raugei, S.; Carloni, P.; Helfand, M. *J. Phys. Chem. B* **2007**, *111*, 2621–2630.
- (8) Klähn, M.; Mathias, G.; Kötting, C.; Nonella, M.; Schlitter, J.; Gerwert, K.; Tavan, P. *J. Phys. Chem. A* **2004**, *108*, 6186–6194.
- (9) Nonella, M.; Mathias, G.; Tavan, P. *J. Phys. Chem. A* **2003**, *107*, 8638–8647.
- (10) Cui, Q.; Karplus, M. *J. Chem. Phys.* **2000**, *112*, 1133–1149.
- (11) Case, D.; III, T. C.; Darden, T.; Gohlke, H.; Luo, R.; Jr, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. *J. Comput. Chem.* **2005**, *26*, 1668–1688.
- (12) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *70*, 926–935.
- (13) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157–1174.
- (14) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648–5652.
- (15) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. *J. Mol. Graphics Modell.* **2006**, *25*, 247–260.
- (16) 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*.
- (17) Berendsen, H. J. C.; Postma, J. P. M.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.
- (18) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. *J. Comput. Chem.* **2005**, *26*, 1701–1718.
- (19) Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471–2474.
- (20) Marx, D.; Hutter, J. Ab initio molecular dynamics: Theory and Implementation. In *Modern Methods and algorithms of quantum chemistry*; Grotendorst, J., Ed.; John von Neumann Institute for Computing: Jülich, 2000; NIC series, Vol. 1, pp 301–349.
- (21) Laio, A.; VandeVondele, J.; Rothlisberger, U. *J. Chem. Phys.* **2002**, *116*, 6941–6947.
- (22) Laio, A.; VandeVondele, J.; Rothlisberger, U. *J. Phys. Chem. B* **2002**, *106*, 7300–7307.
- (23) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098–3100.

- (24) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.
- (25) Troullier, N.; Martins, J. L. *Phys. Rev. B* **1991**, *43*, 1993–2006.
- (26) See, e. g.; Sprik, M.; Hutter, J.; Parrinello, M. *J. Chem. Phys.* **1996**, *105*, 1142. Giannozzi, P. In *Computational Approaches to Novel Condensed Matter Systems, Proceedings of 3rd Gordon Godfrey Workshop on Condensed Physics*; Neilson, D., Das, M. P., Ed.; Plenum: New York, 1995.
- (27) Hutter, J.; Alavi, A.; Deutch, T.; Bernasconi, M.; Goedecker, S.; Marx, D.; Tuckerman, M.; Parrinello, M. CPMD; Copyright IBM Corp 1990–2008, Copyright MPI für Festkörperforschung Stuttgart 1997–2001. <http://www.cpmid.org/> (accessed month day, year).
- (28) Wang, J.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049–1074.
- (29) Marthyna, G. J.; Klein, M. L.; Tuckerman, M. E. *J. Chem. Phys.* **1992**, *97*, 2635.
- (30) T, H.; Dunning, J. *J. Chem. Phys.* **1989**, *90*, 1007–1023.
- (31) Hall, G. G.; Smith, C. M. *Int. J. Quantum Chem.* **1984**, *25*, 881–890.
- (32) Smith, C. M.; Hall, G. G. *Theor. Chim. Acta* **1986**, *69*, 63–69.
- (33) Hay, P. J.; Wadt, W. R. *J. Chem. Phys.* **1985**, *82*, 270–283.
- (34) Gordon, R. G. *J. Chem. Phys.* **1965**, *42*, 3658–3664.
- (35) McQuarrie, D. A. *Statistical Mechanics*; Harper Collins Publisher: New York, 1973.
- (36) Bader, J. S.; Berne, B. J. *J. Chem. Phys.* **1994**, *100*, 8359–8366.
- (37) Herzberg, G. *Molecular Spectra and Molecular Structure*; Krieger Publishing Company: 1991, Vol. 2.
- (38) Porezag, D.; Pederson, M. *Phys. Rev. B* **1996**, *54*, 7830–7836.
- (39) Chaban, G. M.; Gerber, R. B. *Theor. Chem. Acc.* **2008**, *120*, 273–279.
- (40) Martinez, M.; Gaigeot, M. P.; Borgis, D.; Vuilleumier, R. *J. Chem. Phys.* **2005**, *125*, 144106.
- (41) Wheeler, R. A.; Dong, H. *ChemPhysChem* **2003**, *4*, 1227–1230.
- (42) Max, J.-J.; Chapados, C. *J. Chem. Phys.* **2004**, *120*, 6625–6641.
- (43) Zhang, X. K.; Lewars, E. G.; March, R. E.; Parnis, J. M. *J. Phys. Chem.* **1993**, *97*, 4320–4325.
- (44) Dong, J.; Lu, X.; Wei, Y.; Luo, L.; Dunaway-Mariano, D.; Carey, P. R. *Biochemistry* **2003**, *42*, 9482–9490.
- (45) Marzari, N.; Vanderbilt, D. *Phys. Rev. B* **1997**, *56*, 12847–12862.
- (46) Carnevale, V.; Raugei, S.; Micheletti, C.; Carloni, P. *J. Chem. Phys. B* **2007**, *111*, 12327–12332.
- (47) Daněček, P.; Kapitán, J.; Baumruk, V.; Bednářová, L.; Jr., V. K.; Bouř, P. *J. Chem. Phys.* **2007**, *126*, 224513.

CT900131Q

JCTC

Journal of Chemical Theory and Computation

Developing Improved Charge Sets for the Modeling of the KcsA K⁺ Channel Using QM/MM Electrostatic Potentials

Denis Bucher,^{*,†} Leonardo Guidoni,^{†,§} Patrick Maurer,^{†,‡} and Ursula Rothlisberger[†]

Federal Institute of Technology EPFL, Institute of Chemical Sciences and Engineering,
CH-1015 Lausanne, Switzerland

Received April 5, 2009

Abstract: The performance of popular molecular mechanics (MM) force fields in treating problems that involve ion–channel interactions is explored. We have used quantum mechanical/molecular mechanical (QM/MM) calculations to compute the electrostatic potential inside the selectivity filter of the KcsA potassium channel. A comparison is made with the result of classical electrostatic calculations with nonpolarizable MM force fields (AMBER, CHARMM, and GRO-MOS). An effective procedure is proposed to improve force field charges by performing a fit on the electrostatic potential computed along QM/MM simulations, using a dynamical electrostatic potential derived charge set. The optimized charge set is able to reproduce the QM/MM electrostatic potentials along the channel axis within 1–2 kcal/mol, which represents an improvement relative to the corresponding electrostatic potentials obtained with popular MM force fields. By providing quantum mechanical benchmark charges and energies for the KcsA selectivity filter, we hope to facilitate developments toward the modeling of ion channels by providing an objective test as to whether a given implementation of a new, polarizable, model represents a real improvement over existing fixed point charge models.

1. Introduction

Potassium channels form a diverse group of transmembrane proteins that play critical roles in a wide variety of physiological processes, including heart regulation, muscle contraction, neurotransmitter release, neuronal excitability, insulin secretion, cell volume regulation, and cell proliferation.¹ For this reason, an impressive amount of research has been directed toward the understanding of how potassium channels achieve the rapid transport of K⁺ ions without compromising the selectivity of the pore against other ions, such as Na⁺.

Publication of the first high-resolution structure of a potassium channel in 1998² led to a vast number of theoretical studies of pore properties and behavior using all-atoms molecular dynamics (MD) simulations. At the time, MD simulations could not be used for conductance computations due to insufficient simulation time given the frequency of permeation events (~10 ns/ion).^{36,37} However, with the advent of more powerful computers, accurate predictions of the transport rate and ionic specificity have become possible.^{3,4} In the future, such MD simulations may lead, for example, to a better understanding of genetic diseases that originate from the impaired function of ion channels.⁵

The KcsA channel, from bacteria *Streptomyces lividans*, is an ideal prototype system for studies aimed at providing advancement in the description of ion channels, since it is relatively small and it contains the essential structural elements that are shared by all potassium channels.² The conserved TVGYG signature sequence in KcsA forms a narrow constriction in the tetrameric pore called the selectivity filter that is responsible for determining which ions can

* Corresponding author. Phone: +61 (0) 41 466 18 63. Fax: +61 (0) 2 6125 4676. E-mail: dbucher@physics.usyd.edu.au.

[†] Institute of Chemical Sciences and Engineering.

[‡] Present address: Departement de Chimie, Université de Montréal, CP 6128, Succursale Centre-Ville, Montréal, H3C3J7, Canada.

[§] Present address: Università degli Studi dell'Aquila, Dipartimento di Chimica, Ingegneria Chimica e Materiali, Monteluco di Roio 67040, L'Aquila, Italy.

permeate (Figure 1b). The backbone carbonyl oxygen atoms of the TVGYG sequence point toward the center of the pore, forming four distinct ion-binding sites, labeled S_1 – S_4 . Each ion-binding site is composed of eight carbonyl oxygens from two adjacent amino acids (or four carbonyl oxygen atoms and four hydroxyl oxygen atoms from the side chain of Thr-75 in the case of S_4).

The quantitative prediction of the transport rate inside the channels requires a high accuracy of the underlying force field. Since ions are charged species, the accuracy of current computation will be predominantly determined by the accuracy with which the molecular electrostatic potential (MEP) can be computed. From the Arrhenius equation, it follows that an error of ~ 1.4 kcal/mol in the MEP can lead to an error of 1 order of magnitude in the computed current. However, such a level of accuracy is difficult to obtain with current force fields. In the KcsA channel, for instance, a difference of more than 30 kcal/mol was reported between the AMBER electrostatic potential and the quantum mechanical electrostatic potential calculated using HF.⁶ Similarly, in the gramicidin channel, the force field derived potential of mean force for K^+ ions along the channel axis indicates a rather large central barrier that is incompatible with conductance data.^{7,8} These results throw some doubts on the full predictive power of current nonpolarizable force fields for situations where sizable electrostatic fields are present. Considering the importance of the MEP for conduction and selectivity properties of ion channels, it would be highly desirable to evaluate the classical MEP in more cases with direct comparison to a QM reference.

In the following sections, we report QM/MM calculations of the MEP inside the KcsA channel and investigate the ability of an optimized atom-centered pairwise potential to reproduce this QM/MM reference. An effective procedure is proposed that can be used to refine the charge parametrization of a force field in order to reproduce a QM/MM reference potential along a MD trajectory. This method is used to explore problems related to the transferability of the force field, for various occupancy states of the KcsA selectivity filter, in the highly polarized environment of the channel. Finally, free energy perturbation (FEP) calculations are performed in order to measure the K/Na selectivity of the channel with an optimized charge set that includes a description of induced dipole shifts on the carbonyl groups of the filter.

2. Computational Methods

2.1. Model Systems. The system studied here is the KcsA channel structure from *Streptomyces lividans* (PDB code: 1kc4, (Figure 1a, b)). The channel was oriented in such way that the z -axis coincides with the pore axis. Different models (Models 1, 2, A, B, and C) were created to compute the MEP in specific cases. In model 1, the X-ray structure was used. In model 2, water molecules were preserved, and the initial coordinates were obtained from 1 ns of MD simulations of the solvated ion channel, which was taken from our previous work.⁹ In both cases, a QM box of ($20 \times 20 \times 24 \text{ \AA}^3$) was centered on the selectivity filter residues. The

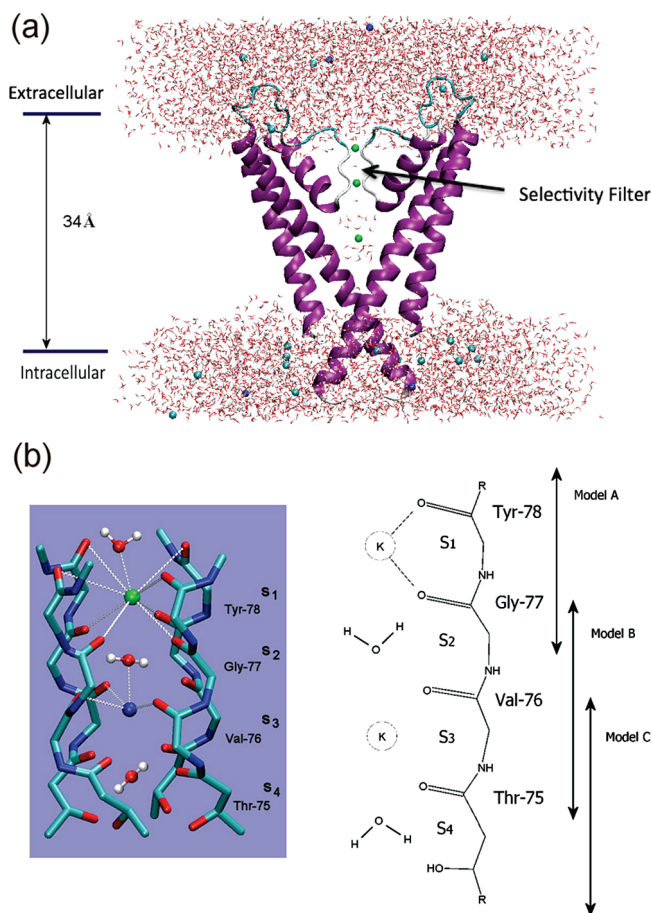


Figure 1. (a) Illustration of the KcsA potassium channel (PDB code 1kc4): only two of the four symmetry equivalent subunits are shown along with the ions in the filter and cavity region. The approximate location of the lipid bilayers is indicated. (b) Selectivity filter with the position of the (S_1 – S_4) binding sites. The occupation (1010) is shown. The entire system (protein + water + membrane) is described classically with the AMBER force field, and a QM box is centered on pairs of residues in the selectivity filter, as indicated by the arrows. Different QM models (A, B, and C) are used to derive an optimized D-RESP force field for the filter atoms.

residues Thr75, Val76, Gly77, and Tyr78 of all four units were simulated using a QM description (208 atoms). The AMBER package, with the parm99 force field,¹⁰ was used for the classical description of the system: the membrane, protein atoms, K^+ ions, water molecules, and solution ions.

In addition, models A, B, and C were created to derive an optimized force field for several filter residues (Figure 1b). The A, B, and C models differ by the filter residues that were included in the QM description. Model A includes as QM the backbone atoms of Tyr78 and Gly77, Tyr78: C_β atom, one K^+ or Na^+ ion, and a water molecule (26 atoms). Model B includes in the QM description: Val76 and one ion (29 atoms). Model C includes in the QM description: Val76:N and Val: C_α atoms, Thr75 residue and one ion (23 atoms).

2.2. Electronic Structure Method. The electronic structure problem in the QM region of the system was solved with DFT. The BLYP functional^{11,12} was used. Core electrons for the K^+ and Na^+ ions were described using

semicore Martins and Trouiller pseudopotentials.¹³ A plane wave cutoff of 80 Ry was used.

A recent claim⁶ that DFT overestimates polarization and charge transfer effects in the KcsA channel could not be confirmed here. The MEP calculations of ref 6 are based on a Mulliken charge analysis in which the Mulliken charge on the K⁺ ion is almost zero. However, the charge on the K⁺ ion is most likely a consequence of the Mulliken charge model, which has been shown to produce unrealistic results for several systems.¹⁴ The two alternative charge models tested here, D-RESP¹⁵ and AIM,¹⁶ both indicated that the charge on the K⁺ ion is close to one. In addition, we find that the lowest energy point for the truncated filter (TVGYG residues only) is ~ 120 kcal/mol at the BLYP level of theory, in agreement with previous HF and B3LYP calculations.^{6,17} The use of a different DFT functional (BP86) did not affect this conclusion.

The influence of the size of the QM box on the MEP was investigated by using a larger QM system ($26 \times 26 \times 24 \text{ \AA}^3$) that included a QM description of the filter residues side chains. In this case, the MEP was only slightly more negative (~ 2 kcal/mol). In addition, small deviations in the orientation of the channel pore, with respect to the *z*-axis, were monitored during the simulations, but they did not affect the MEP significantly.

2.3. QM/MM MD Simulations. QM/MM¹⁸ Car–Parrinello¹⁹ MD simulations were run using the quantum mechanical program CPMD.²⁰ The simulations were performed at a finite temperature (310 K) after classical equilibration in the NTP ensemble. A Nosé–Hoover thermostat^{21,22} was used with a coupling constant of 700 cm^{-1} . The electron fictitious mass was set to 400 au. Three simulations, of 10 ps each, were used to derive an optimized D-RESP charge set for the filter atoms. One additional simulation, of 10 ps, was used to test the ability of the optimized charge set to reproduce the reference QM/MM MEP inside the filter.

2.4. Molecular Electrostatic Potential Calculations. The AMBER/parm99,¹⁰ GROMOS96,²³ and CHARMM 27²⁴ force fields were considered in the classical simulations. MD simulations were run for 5 ns in order to thermally average the MEP over a converged trajectory. A total of 100 snapshots were taken from the trajectory and used to compute the MEP, using everywhere a dielectric constant $\epsilon = 1$. The average fluctuation in the MEP along the channel axis for these snapshots was ~ 5 kcal/mol. In parallel, the MEP was computed for the X-ray structures with a dielectric constant $\epsilon_p = 2$ assigned to the protein to account for thermal vibration occurring in the protein polar groups.²⁵

Classical calculations of the MEP inside the filter were carried out using the APBS package.²⁶ The Poisson equation was solved using the finite difference method. A multigrid method was specified with a coarse mesh of size ($97 \times 106 \times 143 \text{ \AA}^3$) and a fine mesh of size ($20 \times 20 \times 20 \text{ \AA}^3$). A previous lower resolution calculation was used to set the values of the potential at the boundary. The fine mesh was centered on the selectivity filter, and the number of grid points was ($129 \times 129 \times 129$).

2.5. Parameterization of the D-RESP Charge Set. An optimized D-RESP charge set was derived for the backbone

atoms of the KcsA selectivity filter in order to describe the electrostatic potential of the three QM/MM simulations that were carried out for a total time of 30 ps. The simulations were started with models **A**, **B**, and **C**. Charge sets were derived for different occupancies of the S₁–S₄ binding sites and are referred to in the text as charge set: (K⁺), (Na⁺), or vacant (–). The fitting procedure was performed simultaneously on 100 equidistant frames extracted from the ~ 10 ps trajectories. The D-RESP charge sets were optimized so that they reproduced, as well as possible, the electrostatic potential created by the QM/MM system on the nearest classical atoms. Classical atoms within a cutoff distance of 6 \AA from any of the QM atoms (from now on NN atoms) were used as probe sites. The target function *T* to be minimized was

$$T = \sum_{l=1}^L \left(\sum_{j \in \text{NN}} w^v (V_{jl}^{\text{MM}} - V_{jl}^o)^2 + \sum_{j \in \text{NN}} w^A (q_i - q_j^A)^2 \right) \quad (1)$$

where w^v and w^A are weighting factors, and V_{jl}^o is the QM potential on NN atom *j* in snapshot *l*. The second term in eq 1 is used to restraint the value of q_i to the AMBER parm99 charges by a quadratic function. This modification of the original D-RESP scheme¹⁵ is used to derive a charge set for the filter atoms that is as consistent as possible with the AMBER force field. The minimization of *T* was achieved by a least-squares procedure. Values of $w^v = 0.1$ and $w^A = 0.05$ were chosen in the present study since the charges obtained in this way were able to reproduce the QM potential well (standard deviation, $\text{SD}_v \sim 2\%$). By optimizing a different charge set for each frame of a reference QM/MM trajectory, we find that fluctuating charges offer only a small gain in accuracy with respect to fixed point charges ($\text{SD}_v \sim 1\%$ vs $\text{SD}_v \sim 2\%$). However, this conclusion is only valid for the filter binding sites, and charges that respond dynamically to fluctuations in the electric field may prove useful to describe the translocation of ions between binding sites. To overcome the limitation imposed by the fixed point charge model in the proposed scheme, it is possible to derive several charge sets that are optimized for different arrangements of the system. In that case, the energy required to distort the electric clouds and create the dipoles must be described explicitly in the force field, and a self-polarization energy term U_{pol} must be included. Such a term has been called the “missing term” in pair potentials.²⁷

In Figure 2, we illustrate the importance of the missing term for the correct description of the interaction between acetamide and Na⁺ in the gas phase. The QM energy (black line) was computed at the BLYP level of theory with a large basis set (120 Ry). Optimized RESP charges were used to capture the electrostatic energy of the system at different Na–O separation distances (red line with triangle, MK charges²⁸). The default Amber/parm99 partial charges, on the other hand, underestimate the electrostatic energy for this system (green line with diamonds) because the partial charges in biomolecular force fields have been optimized to reproduce the dipoles of carbonyl groups in the average field of proteins. To capture the energy required to create the induced dipoles

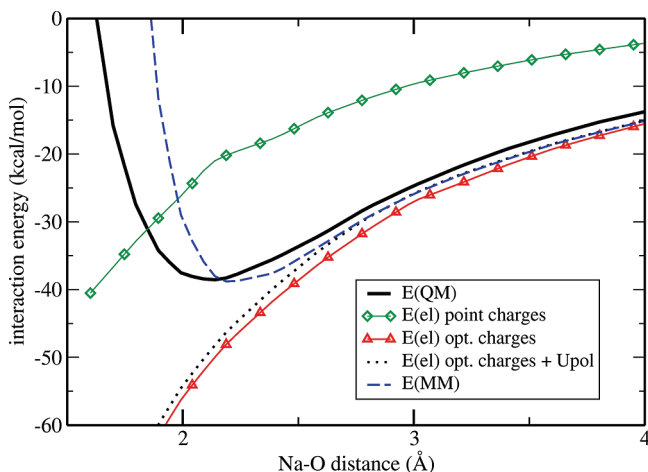


Figure 2. Interaction energy of Na^+ with acetamide in the gas phase. Optimized RESP charges (red line with triangles) are able to describe accurately the electrostatic energy of the system at different Na–O separation distances. The addition of a self-polarization term U_{pol} improves the quality of the fit (dotted line). After the addition of a van der Waals energy term, the total classical energy (blue dashed line) is shown to approximate well the QM energy (black line).

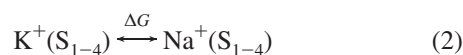
in the RESP scheme, a self-polarization term U_{pol}^{27} can be estimated using the classical formula:

$$U_{\text{pol}} = \frac{1}{2} \mu_{\text{ind}}^2 / \alpha \quad (3)$$

Where μ_{ind} is the ion-induced dipole shift, and $\alpha = 5.41 \text{ \AA}^3$ ²⁹ is the trace of the polarizability tensor for acetamide in the gas phase.

The classical interaction energy E_{MM} (blue dashed line) is obtained by adding the corrected RESP electrostatic energy (dotted line) to the van der Waals energy of the Amber/parm99 force field. A good agreement between the classical and the QM energy is obtained. Further improvement is possible, for example, by describing the repulsive part of the van der Waals potential with a function decaying as R^{-8} instead of R^{-12} .³⁰

2.6. Free Energy Calculations. The energetic of ion selectivity was investigated in MD simulations using both the standard AMBER force field and an optimized D-RESP charge set. MD simulations were performed in the (1010) and (0101) occupancy state of the filter. The thermodynamics of selectivity for each one of the S_1 – S_4 binding sites in the filter was defined as



For each of the FEP computations, the forward and backward free energy perturbation ($\text{K}^+ \leftrightarrow \text{Na}^+$) had values of coupling parameter λ varying from 0 to 1 by increments of 0.05 and was sampled for a total of 1 ns.

3. Results and Discussion

In Figure 3, we report the MEP calculated with classical force fields and the QM/MM reference. Two cases were consid-

ered: (i) a single snapshot in solution (X-ray structure, model 1) and (ii) the thermally averaged MEP in explicit solution computed over a 5 ns trajectory (model 2). In the case of the X-ray structure (PDB code 1k4c), the absolute values for the MEP in the S_2 site were -130 , -136 , -123 , and -128 kcal/mol for AMBER, CHARMM, GROMOS, and the QM/MM reference, respectively (Figure 3b). Model 2 shows how thermal fluctuations influence the MEP profile along the channel axis (Figure 3c). The average fluctuation in the MEP profile along a converged trajectory is ~ 5 kcal/mol. Deviations between the average reference MEP and the average MEP for the three force fields tested are significant (~ 10 kcal/mol) (Figure 3d).

D-RESP charge sets were derived for the selectivity filter atoms. Three charge sets were optimized for various occupancy states of the filter, using 10 ps QM/MM trajectories. The charge parameters are reported in Table 1 together with the charge values of standard force fields. Two tests are performed in order to evaluate the performance of the optimized D-RESP charge sets: (i) we measure the ability of the charge set to reproduce the QM/MM reference MEP along the z -axis of the channel, and (ii) we evaluate the ability of the charge set to reproduce the reference electrostatic interaction of the filter atoms with the surrounding protein environment (NN atoms) for a new trajectory. In both tests, deviations between the D-RESP and QM/MM electrostatic potential are found to be smaller than 5%.

The optimized D-RESP charge sets were able to reproduce the reference MEP inside the filter within $\sim 2\%$ (Figure 3b,d), and the interaction between the NN atoms and the filter atoms was improved (Table 2). In the case of the AMBER force field, on the other hand, typical deviations between the QM/MM reference and the classical MEP are on the order of $\sim 20\%$. The transferability of the force field parameters to different occupancy states of the filter was also improved. For example, an optimized charge set derived in the presence of Na^+ is also found to perform well in the presence of K^+ ions (Table 2).

The magnitude of ion-induced polarization shift in the carbonyl groups of the filter can be estimated from the optimized charge sets. The potassium-induced shift in the carbonyl dipoles is found to be ~ 0.3 D, in good agreement with previous estimates from QM^{31,32} and QM/MM³³ calculations. The sodium-induced shift is 0.7 D, which is slightly lower than a previous QM/MM estimate of ~ 1.5 D³³ based on a Wannier functions analysis. However, it should be noted that the optimized point charge model can only capture sodium-induced shifts that occur along the carbonyl axis, and therefore, it may slightly underestimate the polarization effects. The amount of charge transferred from the carbonyl groups to the K^+ ions ($\sim 0.05 e$) is in good agreement with estimates from QM,^{17,31,32,34} and QM/MM calculations³³ (values range between 0.05 and 0.15 e). These encouraging results suggest that the optimized D-RESP charges offer a mean to capture ion-induced polarization shift on the carbonyl groups of the KcsA selectivity filter.

An optimized charge set was used to compute the K/Na selectivity of the filter binding sites. Noskov et al. have previously shown that coordinating ligands with very large

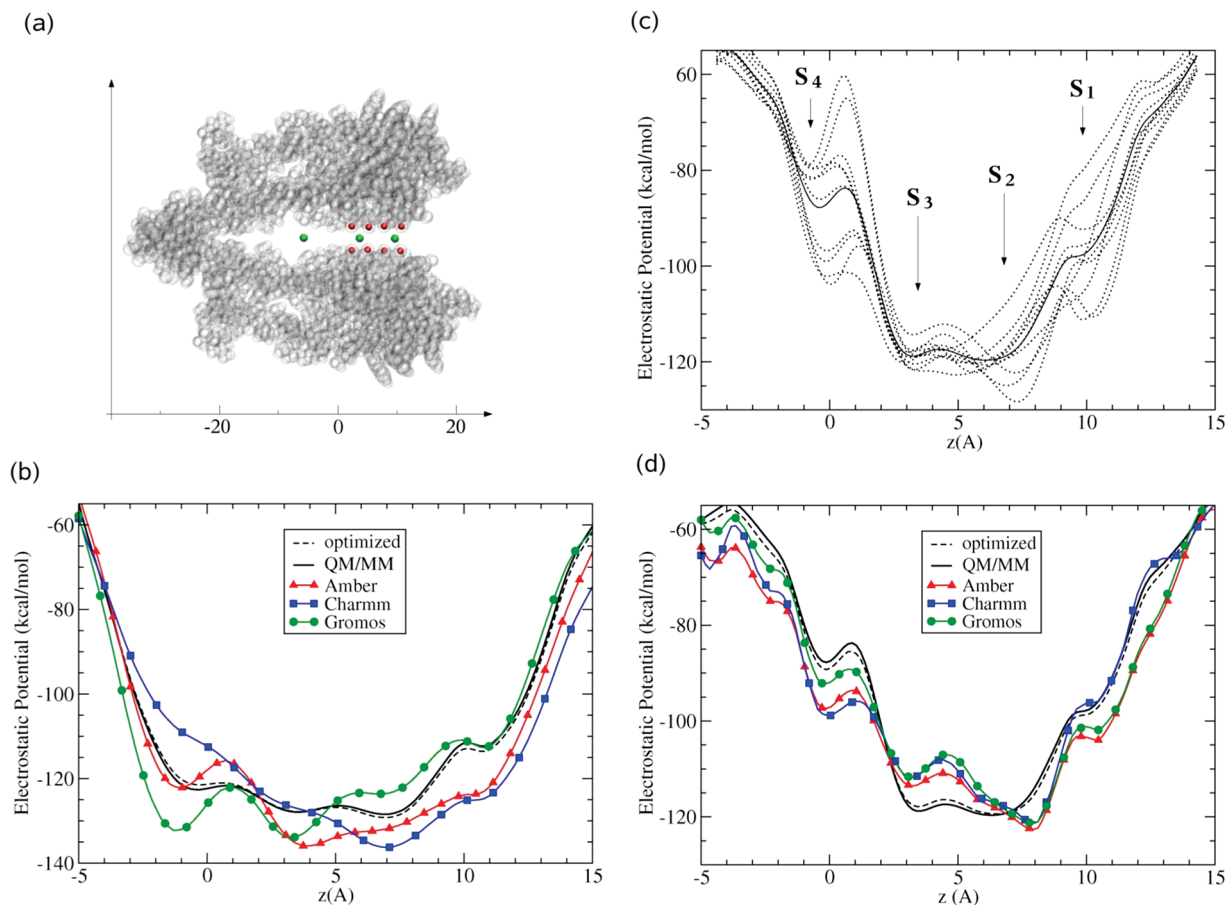


Figure 3. (a) Schematic diagram of the KcsA potassium channel. Carbonyl oxygens that form the S_1 – S_4 potassium binding sites are displayed as red spheres. K^+ ions are displayed as green spheres. (b) Electrostatic potential inside the KcsA selectivity filter in the case of the X-ray structure (model 1). Comparison between force fields and QM/MM calculations. (c) Typical fluctuations in the electrostatic potential shown for 10 calculations taken from a 1 ns trajectory (model 2). The positions of the (S_1 – S_4) binding sites are indicated by arrows. (d) Electrostatic potential inside the KcsA selectivity filter. The force fields electrostatic potential and the QM/MM electrostatic potential are compared after the field created by the water has been thermally averaged.

Table 1. Average Charges for the Filter Backbone Atoms^a

atoms	AMBER 94/99 ¹⁰	GROMOS96 ²³	CHARMM22/27 ²⁴	D-RESP (K^+)	D-RESP (Na^+)	D-RESP (–)
C	0.597	0.380	0.510	0.607 (0.019)	0.603 (0.019)	0.546 (0.022)
O	–0.569	–0.380	–0.510	–0.532 (0.020)	–0.656 (0.019)	–0.500 (0.021)
N	–0.416	–0.280	–0.470	–0.394 (0.016)	–0.413 (0.016)	–0.393 (0.017)
CA	0.035	0.000	0.000	–0.027 (0.014)	0.029 (0.016)	–0.021 (0.018)
HA	0.048	0	0.058	0.055 (0.012)	0.051 (0.010)	0.058 (0.013)
HN	0.272	0.280	0.310	0.261 (0.004)	0.262 (0.004)	0.254 (0.004)
(C + O + N + CA + HA)	–0.305	–0.280	–0.412	–0.298	–0.388	–0.316
Na^+	1.000	1.000	1.000	–	0.948 (0.016)	–
K^+	1.000	1.000	1.000	0.962 (0.017)	–	–
amide plane dipole (Debye)	4.3	3.7	4.2	3.9	4.9	3.8
C=O dipole (Debye)	3.4	2.2	3.0	3.3	3.7	3.0

^a The occupancy of the S_1 binding site in the reference QM/MM calculation is indicated by (K^+), (Na^+), or (–). Average fluctuations in the D-RESP charge set are indicated in parentheses.

dipoles (~ 7 D)³⁵ are required in order to fully reverse the K/Na selectivity of the binding sites. Here, a protocol similar to ref 35 was used in order to perform FEP calculations and compute the K/Na selectivity of the filter binding sites with an optimized charge set. With the default partial charges of AMBER/parm99, the selectivity per binding site was 2.6, 5.8, 3.8, and 0.3 kcal/mol for the S_1 – S_4 binding sites, in good agreement with CHARMM values of $\Delta G = (2.6, 5.3, 1.8, \text{ and } -1.2 \text{ kcal/mol})$ ³⁵ and

with experimental estimates of $\Delta G = 5$ – 6 kcal/mol for the free energy of selectivity of the potassium channel.^{36,37} With the optimized D-RESP(Na) charge set, the average K/Na selectivity was reduced on average by ~ 0.9 kcal to $\Delta G = (0.7, 4.8, 2.1, \text{ and } 0.3 \text{ kcal/mol})$. Therefore, the use of an optimized charge set in the present study did not alter previous conclusions about the K/Na selectivity of the filter binding sites, which is indicated by FEP calculations based on nonpolarizable force fields.^{35,38–41}

Table 2. Transferability of the D-RESP Charge Set Parameters^a

occupancy state	AMBER	D-RESP (-)	D-RESP (K ⁺)	D-RESP (Na ⁺)
KWKW	18.2	8.3	1.7	7.3
NaWKW	16.9	12.9	7.5	3.9
-W-W	93.5	7.5	39.6	62.5

^a The relative deviations [%] are reported between the QM/MM (selectivity filter) electrostatic potential on neighboring MM atoms and the same potential calculated with the classical charge sets for 10 ps trajectories. The D-RESP charge sets were optimized in the presence of K⁺ and Na⁺ ions and in the empty filter.

4. Concluding Remarks

At the present time, atomistic conductance computations in an ion channel remain very challenging because of both the long sampling time involved and the very high accuracy required in the computation of the electrostatic potential along the channel axis. In this paper, we have discussed an effective procedure to improve the accuracy of MEP calculations by using a perturbative approach to optimize a force field charge parametrization along QM/MM trajectories. The method is potentially useful to improve the transferability of standard nonpolarizable force fields to the highly polarized environments of the channels. By deriving the parameters for ion configurations of the filter that are frequently sampled, the D-RESP optimization scheme is ideally suited for MD simulations. In some cases, the method can be extended by using a “force-matching” algorithm to derive other force field parameters of interests.⁴² Such a method has already been applied successfully to describe DNA–protein interactions and organometallic compounds.⁴³ More generally, we hope that the benchmark charges and energies presented here for the KcsA channel will facilitate developments toward the modeling of ion channels by providing an objective test as to whether a given implementation of a polarizable model represents a real improvement over existing fixed charge models.

Briefs

Developing improved charge sets for the modeling of a potassium channel using QM/MM electrostatic potentials.

References

- Hille, B. *Ion Channels of Excitable Membranes*, 3rd ed.; Sinauer Associates: Sunderland, MA, 2001.
- Doyle, D. A.; Morais Cabral, J.; Pfuetzner, R. A.; Kuo, A.; Gulbis, J. M.; Cohen, S. L.; Chait, B. T.; MacKinnon, R. *Science* **1998**, *280*, 69–77.
- Sotomayor, M.; Vasquez, V.; Perozo, E.; Schulten, K. *Biophys. J.* **2007**, *92*, 2771–2784.
- Aksimentiev, A.; Schulten, K. *Biophys. J.* **2005**, *88*, 3745–3761.
- Ashcroft, F. M. *Ion Channels and Disease: Channelopathies*; Academic Press: San Diego, CA, 2000.
- Bliznyuk, A. A.; Rendell, A. P. *J. Phys. Chem. B.* **2004**, *108*, 13866–13873.
- Bastug, T.; Kuyucak, S. *Biophys. J.* **2006**, *90*, 3941–3950.
- Edwards, S.; Corry, B.; Kuyucak, S.; Chung, S. H. *Biophys. J.* **2002**, *83*, 1348–1360.
- Bucher, D.; Guidoni, L.; Rothlisberger, U. *Biophys. J.* **2007**, *93*, 2315–2324.
- Case D. A.; Pearlman D. A.; Caldwell J. W.; Cheatham T. E., III; Ross, W. S.; Simmerling, C. L.; Darden, T. L.; Marz, K. M.; Stanton, R. V.; Cheng, A. L.; Vincent, J. J.; Crowley, M.; Tsui, V.; Radmer, R. J.; Duan, Y.; Pitera, J. Massova, I.; Seibel, G. L.; Digh, U. C.; Winer, P. K.; Kollman, P. A. *AMBER*, Version 6.0; University of California: San Francisco, CA, 1999.
- Lee, C. T.; Yang, W. T.; Parr, R. G. *Phys. Rev. B: Condesn. Matter Mater. Phys.* **1998**, *37*, 785–789.
- Becke, A. D. *Phys. Rev. A: At., Mol., Opt. Phys.* **1988**, *38*, 3098–3100.
- Trouiller, N.; Martins, J. L. *Phys. Rev. B: Condesn. Matter Mater. Phys.* **1991**, *43*, 1993–2006.
- Guerra, C. F.; Handgraaf, J. W.; Baerends, E. J.; Bickelhaupt, F. M. *J. Comput. Chem.* **2004**, *25*, 189–210.
- Laio, A.; VandeVondele, J.; Rothlisberger, U. *J. Chem. Phys. B* **2002**, *106*, 7300–7307.
- Bader, R. F. W. *Atoms in Molecules*; Oxford University Press: New York, 1994.
- Huetz, P.; Boiteux, C.; Compoin, M.; Ramseyer, C.; Girardet, C. *J. Chem. Phys.* **2006**, *124*, 44703.
- Laio, A.; VandeVondele, J.; Rothlisberger, U. *J. Chem. Phys.* **2002**, *116*, 6941–6947.
- Car, R.; Parrinello, M. *Phys. Rev. Lett.* **1985**, *55*, 2471–2474.
- Hutter, J.; Alavi, A.; Deutch, T.; Bernasconi, M.; Goedecker, S.; Marx, D.; Tuckerman, M.; Parrinello, M., *Car Parrinello Molecular Dynamics*; International Business Machines Corporation U.S. and the Max-Planck-Institut fur Festkorperforschung: Stuttgart, Germany, 1995–1999.
- Nose, S. *J. Chem. Phys.* **1984**, *81*, 511–519.
- Hoover, W. G. *Phys. Rev. A: At., Mol., Opt. Phys.* **1985**, *31*.
- Huenenberger, I. G.; Tironi, A. E.; Mark, S. R.; Billeter, J.; Fennen, A. E.; Torda, T.; Huber, P.; Krueger van Gunsteren, W. F. *J. Phys. Chem. A* **1999**, *103*, 3596–3607.
- Brooks, B. R.; Brucoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S. *J. Comput. Chem.* **1983**, *4*, 187–217.
- Berneche, S.; Roux, B. *Biophys. J.* **2002**, *82*, 772–780.
- Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10037–10041.
- Berendsen, H. J. C. *J. Phys. Chem.* **1987**, *91*, 6269–6271.
- Besler, B. H.; Merz, K. M., Jr.; Kollman, P. A. *J. Comput. Chem.* **1990**, *11*, 431.
- Patel, S.; MacKerell, A. D., Jr.; Brooks, C. L., III. *J. Comput. Chem.* **2004**, *25*, 1504–1514.
- Roux, B.; Karplus, M. *J. Comput. Chem.* **1995**, *16*, 690–704.
- Bliznyuk, A. A.; Rendell, A. P.; Allen, T. W.; Chung, S. H. *J. Phys. Chem. B.* **2001**, *105*, 12674.
- Guidoni, L.; Carloni, P. *Biochim. Biophys. Acta* **2002**, *1563*, 1–6.

- (33) Bucher, D.; Guidoni, L.; Raugei, S.; Del Perraro, M.; Carloni, P.; Rothlisberger, U. *J. Biophys. Chem.* **2006**, *1241*, 292–301.
- (34) Varma, S.; Sabo, D.; Rempe, S. B. *J. Mol. Biol.* **2008**, *376*, 13–22.
- (35) Noskov, S. Y.; Berneche, S.; Roux, B. *Nature*. **2004**, *431*, 830–834.
- (36) LeMasurier, M.; Heginbotham, L.; Miller, C. *J. Gen. Physiol.* **2001**, *118*, 304–314.
- (37) Nimigean, C. M.; Miller, C. *J. Gen. Physiol.* **2002**, *120*, 323–325.
- (38) Luzhkov, V. B.; Aqvist, J. *Biochim. Biophys. Acta* **2001**, *1548*, 194–202.
- (39) Allen, T. W.; Bliznyuk, A.; Rendell, A. P.; Kuyucak, S.; Chung, S. H. *J. Chem. Phys.* **2000**, *112*, 8191–8204.
- (40) Thomas, M.; Jayatilaka, D.; Corry, B. *Biophys. J.* **2007**, *93*, 2635–2643.
- (41) Berneche, S.; Roux, B. *Nature*. **2001**, *414*, 73–77.
- (42) Maurer, P.; Laio, A.; Hugosson, H. W.; Colombo, M. C.; Rothlisberger, U. *J. Chem. Theory Comput.* **2007**, *3* (2), 628–639.
- (43) Spiegel, K.; Magistrato, A.; Maurer, P.; Ruggerone, P.; Rothlisberger, U.; Carloni, P.; Reedijk, J.; Klein, M. L. *J. Comput. Chem.* **2008**, *29*, 38–49.

CT9001619

Effects of Hydration on the Conformational Energy Landscape of the Pentapeptide Met-Enkephalin

L. Ramya and N. Gautham*

*Centre of Advanced Study in Crystallography and Biophysics, University of Madras,
Chennai, 600025, India*

Received January 5, 2009

Abstract: We report here a study of the conformational energy landscape of the pentapeptide Met-enkephalin in the presence of explicit solvent (water) molecules. A sample of 1 500 low-energy structures of this molecule was generated using the mutually orthogonal Latin squares (MOLS) technique with the CHARMM22 force field. This technique, developed in our laboratory, allows us to sample the conformational space of a molecule in an unbiased and exhaustive manner. The study shows that inclusion of explicit solvation is important to correctly model the conformational behavior of the molecule. Structures modeled in the presence of water molecules are far more similar to the experimental structures than when the water molecules are excluded. The results also indicate that the pentapeptide Met-enkephalin prefers extended structures in an aqueous environment, as against tightly folded structures in the absence of water. Thus, the biologically relevant structure, when the molecule is not bound to the receptor, is probably the extended structure, as seen in the crystallographic and NMR experiments, rather than the GEM structures calculated by various workers.

Introduction

Enkephalins are endogenous opioid neuropeptides that have morphine-like activity.¹ One of these, Met-enkephalin, is an inhibitory neurotransmitter, and its action results in decreased pain sensitivity. This pentapeptide takes on different conformations when it binds to the three subclasses of opioid receptors² termed μ , δ , and κ based on selectivity of different agonists for different physiological responses.³ It has been observed that a β -bend folded conformation of the neuropeptide binds to the μ receptor, while an extended conformation binds to the δ receptor.⁴ The peptide has a flexible conformation, which is strongly influenced by the local environment. There is evidence that the bound conformations are different from the crystal⁵ and solution structures.⁶

In such contexts, computational studies on the peptides become important. The enkephalins, and in particular Met-enkephalin, have been subjected to numerous such studies.^{7,8} Many theoretical conformational studies of Met-enkephalin have been carried out using ECEPP, AMBER, and CHARMM force fields.⁷ The low-energy conformations of Met-en-

kephalin calculated in the absence of explicit solvent effects have been determined based on the ECEPP/2 and ECEPP/3 force fields.⁸ Eisenmenger and Hansmann⁸ have studied the variation of the energy landscape of Met-enkephalin by changing the force field from ECEPP/2 to ECEPP/3. When the force field is changed, the low-energy conformation of the peptide changes, as reflected in the energy landscape, but most of the physical results, such as the thermodynamic properties, are not affected. In the calculations using the ECEPP/2 potential,⁹ two different characteristic temperatures of the peptide were determined. These were a collapse temperature T_θ , corresponding to coil-globular transition, and a folding temperature T_f , corresponding to one of the many local minima. Zhan and associates¹⁰ studied the conformations of Met-enkephalin using the ECEPP/2 and ECEPP/3 force fields. With the ECEPP/3 force field, they reported a new global minimum, different from the one located by Isogai et al.¹¹ All the computations described above were carried out without explicitly specifying the solvent water.

Many other studies have been carried out with implicit solvent.^{12–16} Klepsis and Floudas¹² calculated the free energy of the peptide, both with and without solvent, using the α BB

* Corresponding author. E-mail: n_gautham@hotmail.com.

global optimization algorithm. Their results indicated that both cases produced large numbers of low-energy local minima. For the consideration of hydration effects, they adopted a model based on the solvent accessible surface area as well as the solvent accessible volume. Significant differences were observed in the two models, and this led to the conclusion that appropriate model selection is essential for computing hydrated peptide structures. Caracci¹³ studied the thermodynamic properties of the peptide in terms of its ionization state and the solvent environment using Monte Carlo simulated annealing. Again the hydration effects were modeled in terms of the solvent accessible surface area.¹⁴ This analysis supported a mechanism of interconversion between the extended conformer and three bent conformers. The relative free energies of the dominant conformers are not very large, particularly in the solvation calculations. The ensemble entropy is similar in magnitude to the experimental entropy of the unfolding peptide. Shen and Freed¹⁵ performed molecular dynamics simulations for Met-enkephalin with both explicit and implicit solvent models. They observed that the two simulations gave similar results. Simulations of Met-enkephalin have also been carried out using explicit and united atom potentials.¹⁶ The results indicated that the two force fields produced significantly different peptide dynamical properties.

In an earlier report from our laboratory, we explored the conformational space of Met-enkephalin in an exhaustive manner using a technique we had developed called the mutually orthogonal Latin squares technique.¹⁷ The ECEPP/3 potential function was used in the calculations,¹⁸ which were performed with no specific model for water and solvent interactions. It was assumed that the parametrization of the force field based on experimental data automatically incorporated solvent effects. The results of these studies suggested the following: (a) The backbone of the peptide exists as a mixed form of both folded and unfolded structures; (b) The fraction of folded structures is low compared to extended conformations; and (c) Some of the low-energy structures correspond to the fully extended conformation, similar to the crystal structure.

In this paper, we present the results of our calculations of the conformational energy landscape of Met-enkephalin in the presence of explicit water molecules. We perform the calculations using the same mean-field method with MOLS sampling, as in the case described above. However, we have now further enhanced the technique to include explicit water molecules and their interactions. In the Methods and Materials section, we describe the technique, along with details of how it has been modified to include explicit water molecules. The results of the computations are then discussed in the context to the other studies that have been carried out on Met-enkephalin and other peptides, both with and without explicit solvent.

Methods and Materials

The mean-field technique (MFT) with MOLS sampling, used to exhaustively explore the conformational space of oligopeptides, has been extensively described elsewhere.^{17–19} A detailed explanation of the MOLS method is also available

at http://www.unom.ac.in/Gautham_mols.pdf. For completeness, we repeat the main points of the description here.

MFT^{20,21} requires the conformational space to be divided into a number of subspaces, each of which may have a number of states. The effective potential, obtained by setting a particular subspace to a particular state, is then obtained by taking a probability-weighted average of the pairwise interactions between that state of that subspace and all the states of all other subspaces. The probabilities, initially set to be the same for all the states of a given subspace, are then recalculated as the Boltzmann function of the effective potentials. These new probabilities are then used in the next cycle to recalculate the effective potentials. The iterative cycles are repeated until the probabilities (and the effective potentials) converge. The states with the highest probabilities after convergence define the optimum in the conformational space. This procedure has been applied, for example, to determine the optimum side-chain conformations of a protein, given the backbone fold.²¹ In this case, the conformation of each side chain constitutes a subspace and the various possible conformers for each constitute its states. The effective potential of setting a particular side-chain to a particular conformer is calculated as the probability-weighted sum of the pairwise interactions between that side chain and each of the conformers of each of the other side-chains in the protein. Upon convergence, the most probable conformers define the structures of the side-chains of the protein.

A crucial factor in the success of such a straightforward application of MFT, in this case, is the relative independence of the side chain structures from each other. Each term in the sum adding up to the effective potential depends only on the conformation of the particular pair of side chains and is independent of the other residues. Thus, the interaction of the residue A with B is considered to depend only on the side chain structures of these two residues and to be independent of the others, an assumption which turns out to be largely valid.

When the method is attempted to be carried over to calculations of the protein backbone conformation, however, a similar assumption of independence does not hold. One may consider each of the backbone torsion angles to be a subspace and the values that are possible for it (for example, any value from 0 to 360° in steps of, say, 10°) as the states accessible to it. However, it is clear that, in this case, the effective potential due to setting one of the torsion angles to one of the values cannot simply be calculated as a weighted sum of pairwise interactions. This is because the interactions that may occur due to the settings of a given pair of torsion angles depend also on the settings of all the intermediate torsions. A rigorous calculation of the effective potential would then require the consideration of all settings of all torsion angles, leading to combinatorial explosion.

To avoid this exponential increase in computation time, we use mutually orthogonal Latin squares (MOLS) sampling to identify a small sample of the conformational space, which is then used to compute the effective potential. MOLS sampling is used in the design of experiments in, for example, agriculture or clinical trials.²² The common feature in these studies is the large, multivariable experimental space. MOLS

A	B	C
B	C	A
C	A	B

Figure 1. A simple Latin square is of order 3 (adapted from Cochran and Cox²⁴).

A α	B β	C γ
B γ	C α	A β
C β	A γ	B α

Figure 2. Two mutually orthogonal Latin squares (MOLS) of order 3. The Latin letters A, B, and C are symbols of the first Latin square. The Greek letters α , β , and γ are symbols of the second Latin square, which is orthogonal to the first Latin square (adapted from Fisher²⁵).

is used to identify a small representative sample of this space, and the experiments or calculations are carried out only for this sample, rather than for the entire space. A Latin square of order n is an arrangement of n symbols in a $n \times n$ square array in such a way that each symbol occurs exactly once in every row and column.²³ A simple Latin square of order 3 is shown in Figure 1, where the Latin letters (A, B, and C) are used as symbols to construct the square. Two Latin squares of order n are called orthogonal if, when one is superimposed on the other, all n^2 possible combinations of the two sets of n symbols are present, and each symbol of the one square occurs once, and only once, with each symbol of another square. Figure 2 shows a pair of orthogonal Latin squares involving two sets of three elements each. It was constructed by the superposition of two Latin squares, one square consisting of Latin letters (A, B, and C) and the other consisting of Greek letters (α , β , and γ). It may be seen from Figure 2 that each subsquare consists of one Latin letter and one Greek letter, and each combination of Greek and Latin letters appears only once. In other words the symbols are arranged such that all n^2 possible pairs appear in the array. The concept can be extended to N mutually orthogonal Latin squares (MOLS), every pair of which is orthogonal.²³ Thus, in a set of three MOLS (i.e., $n = 3$), the first square would be orthogonal to both the second and third squares, and the second square would also be orthogonal to the third. It has been shown that if n is a prime number or any integer power of a prime, then it is possible to construct a complete set of $n - 1$ mutually orthogonal Latin squares of order n .²⁶

A single Latin square of order n allows one to investigate the effect of setting three subspaces (or factors in the language of experimental design), each at n different states (or values) in only n^2 trials, instead of n^3 trials that would

be required for a complete and exhaustive sampling. A set of m mutually orthogonal Latin squares of order n allows one to investigate the effect of setting m factors each to different n values, again in only n^2 (instead of n^m) trials. It may be seen from these descriptions, that a distinctive feature of using MOLS to select the sampled points is the occurrence in the sample of every possible pairwise setting of all the states of all subspaces.

In applying this technique to the exploration of conformational space, including the portion mapped out by the backbone torsion angles of a peptide chain, we identify the subspaces with the torsion angles, there being m such angles, and the states as the values that the torsion angles can take, there being n such values accessible to each torsion angle. Out of the total of n^m points available in the conformational space, we select n^2 points using MOLS and calculate the potential energy corresponding to each of these conformations. To calculate the MFT effective potential due to setting a torsion angle to a specific value, we take the Boltzmann-weighted average of the potential energies at all those n points in the MOLS grid in which that value of that torsion angle occurs. We note that, as described above, this ensures the *pairwise sampling* of that value of the torsion angle with every value of every other torsion angle. This procedure may be repeated for all values of all torsion angles, using the same MOLS grid specifying the n^2 conformations. At the end of this set of calculations, we obtain the effective potentials corresponding to every setting of each torsion angle. In the normal MFT procedure, these potentials are used to recalculate the probabilities, which are, in turn, used to recalculate the effective potentials in iterative cycles until convergence.²¹ In the present method using MOLS sampling, however, this is not possible since the effective potential is itself evaluated as a Boltzmann-weighted average of the potentials calculated in the MOLS grid. Thus, the effective potentials at the end of one round of calculations described above are used directly as surrogates for the final probabilities. For each torsion angle, the value with the lowest effective potential is considered the most probable one, and the set of most probable values for all the torsion angles defines the optimum conformation of the peptide.

The crucial assumption underlying the MFT procedure, in particular when used with MOLS sampling, is regarding the degree of independence of the variables (the torsion angles). The procedures work best when the variables are completely independent of each other. However, in that case, they are not required, and a few simple one-dimensional searches would quickly identify the optimum. Backbone torsion angles are highly interdependent, and the procedure outlined above may not lead always to a single global optimum. A second assumption implicit in the procedures is that such a single, global optimum exists on the potential energy landscape. In fact, as many studies have shown,⁸ most definitions of the force field operating in proteins and peptides lead to several optima (minima) in conformational space, with approximately equal energy values. We have established¹⁹ that the MOLS sampling procedure offers a way of exploring the entire space in an exhaustive manner (at least for peptides up to about 10 residues in size) and

identifying all the optima. As stated above, the construction of one MOLS grid of order n (and the consequent calculation) leads to one optimal or low-energy conformation. Repeating this procedure by using a different set of MOLS leads to another. There are $(n!)^m$ different ways of constructing the MOLS grid²⁷ and, in principle, the calculations may be repeated as many times to identify all the low-energy conformations. Of course, this does not mean that there are $(n!)^m$ low-energy conformations for a peptide. Most of the conformations generated in subsequent cycles are the same as one generated previously. We, therefore, use a clustering algorithm²⁸ to cluster the structures together and identify the unique ones. In practice, we have found that for a pentapeptide it is sufficient to generate 1 500 conformations to identify all its 50 or so low-energy conformations.

We have shown¹⁷ that the sampling of the conformational space of Met-enkephalin carried out by this procedure repeated over a few hundred cycles is exhaustive, and all low-energy structures are identified. The library of structures generated in this manner includes the global energy minima of other simulations as well as the structures obtained by experiments. It also includes some well-folded structures that were not earlier identified by any other experimental or computational method.

In the above reports, we had used the ECEPP/3 semiempirical force field, without including any explicit water molecules. Solvent interactions were deemed to have been incorporated in the force field parametrization. In the present paper, we have extended the method to include explicit water molecules in the calculations. This has been done in the following manner. First, explicit water molecules were added to each of the n^2 MOLS-defined conformations using the program VMD.²⁹ The peptide was solvated with a water box such that there was a layer of water 5 Å in each direction from the ends of the molecule in each of the three directions. Thus the structure, even if fully extended, was completely immersed in the water. Next, the conformational energy of the molecule was calculated using the CHARMM22 potential.³⁰ The force field included the torsion angle energy and electrostatic, van der Waals, and solvent interactions. Hydrogen bond energies are implicitly included in the CHARMM potential energy function³⁰ through nonbonded electrostatic and van der Waals interactions. The TIP3P model³¹ of water molecules, as available in the CHARMM force field, was used to calculate the solvent interactions. The n^2 potential energies calculated were then used in the rest of the MOLS procedure to obtain the effective energies for each value of every conformational variable and, subsequently, to identify the optimum conformation, now in the presence of water. The optimal conformation obtained was minimized for 1 000 steps by the conjugate gradient method using the NAMD³² minimization routine, once again with explicit water molecules.

Using this protocol, 1500 structures were generated for the pentapeptide. These 1 500 minimized structures were clustered using a hierarchical clustering method,²⁸ with an rmsd cutoff of 1.0 Å for the backbone atoms, i.e., two structures with backbone atom rmsd of 1.0 Å or less were considered to be the same structure and only one of them was retained in the list of unique structures. The procedure

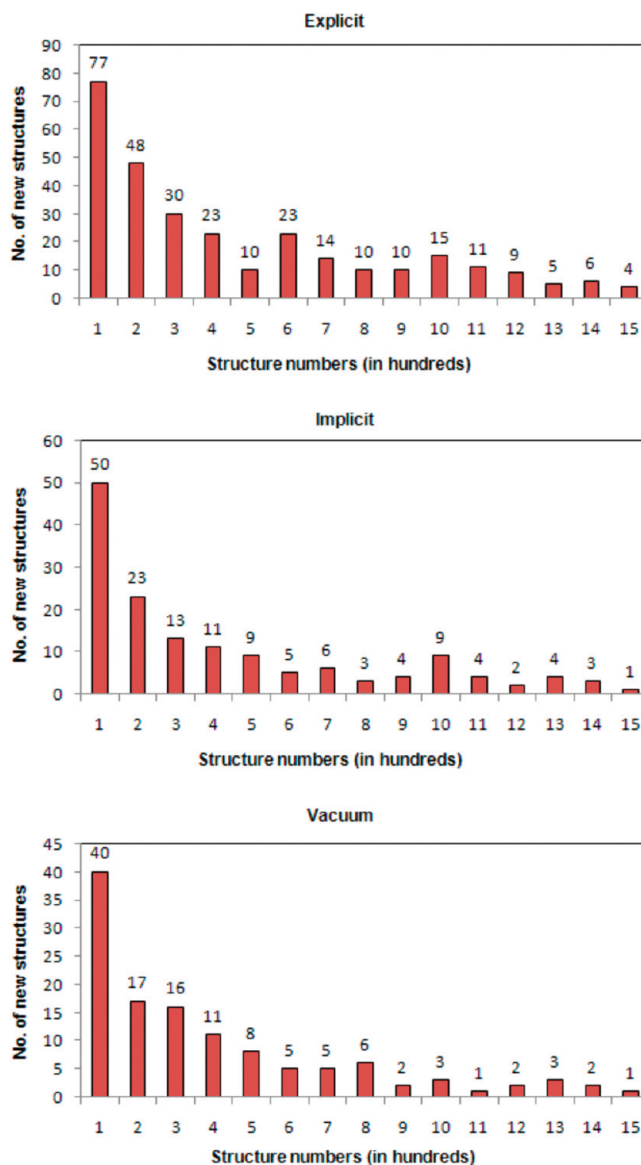


Figure 3. Number of unique structures obtained in every set of 100 structures generated by the MOLS method.

yielded 295 unique structures. Figure 3 shows the number of unique structures obtained in every set of 100 structures generated by the MOLS method. The figure indicates exhaustive sampling of the conformational space of the solvated peptide, since very few new structures are discovered after about 1 200 MOLS structures. The energies used for subsequent comparisons and analyses were those obtained from the NAMD minimization routine, which includes the bond length and angle, dihedral angle, and improper and nonbonded energies. The energies of interaction between the water molecules and the peptide atoms were included in the subsequent analyses but not the intramolecular energies of interaction among the water molecules. This entire procedure was repeated for another set of 1 500 structures; in this case, however, water molecules were not included, thereby yielding a set of structures calculated in vacuum. The set of structures generated by these calculations were clustered using the hierarchical algorithm and resulted in 122 clusters. The computations required 45 min to generate one explicit

solvated structure on an AMD Opteron processor at 2.2 GHz and 7 min on the same platform for each nonsolvated structure.

In order to compare the effects of implicit solvent in the structure prediction, we used the implicit solvent model analytical continuum electrostatic potential³³ (ACE) along with the CHARMM22 force field. (Since NAMD does not support the implicit solvent model, in these calculations we used the program CHARMM in its place, i.e., to calculate the energies and the final minimization step after identifying the low-energy conformations using the MOLS procedure identical to the one above.) As in the case with explicit solvent, 1 500 conformations were generated. These structures were clustered using the hierarchical clustering method with an rmsd cutoff of 1.0 Å for the backbone atoms. This yielded 147 unique structures. The energies used for subsequent comparisons and analyses were those obtained from the CHARMM minimization routine. The computations required 4 min to generate one implicit solvated structure on an AMD Opteron processor at 2.2 GHz.

Results and Discussion

We first discuss the effect of explicit solvation on the global minimum-energy structures, followed by the analyses of the structural motifs in the generated structures. Next, we describe the patterns and distribution of hydrogen bonds. Finally, we describe the effect of the solvation on the energy landscapes.

A. Minimum Energy Structures. Figure 4 compares the minimum-energy structures obtained by the MOLS method with and without solvation. The two structures are quite different from each other, with a rmsd of 3.32 Å on least-squares superposition. In the solvated structure, the side chain of methionine faces the aromatic ring of tyrosine. In the nonsolvated structure, the methionine side chain faces the aromatic ring of phenylalanine. Table 1 gives the backbone rmsd on the superposition of both solvated and nonsolvated lowest-energy structures with the global energy minimum (GEM) structure of Scheraga and co-workers,¹¹ two NMR structures,³⁴ and the X-ray crystal structure.⁵ The table shows that, first, the nonsolvated structure is closer to the GEM than the solvated one. This is as expected, since the GEM structure was computed in the absence of explicit solvent. Second, when superposed on all three experimental structures, the explicit solvated structure has a smaller rmsd value. Again, this matches with expectations, since the experiments were performed on the solvated molecule.

As noted in our previous computations,¹⁸ of the structures generated, the lowest-energy structure is only very rarely the best-sampled one, i.e., one with the lowest rmsd as compared to either the GEM or the experimental structures. This is true, in the present case too, and Table 2 gives the rmsd and the energy values of the best sampled structures with and without solvent (Figure 5). The rmsd values indicate that the solvated best sampled structures are only marginally closer to the experimental structures and the GEM than are the nonsolvated best sampled structures.

Table 3 gives the results of the comparison of all the MOLS low-energy structures generated, with and without

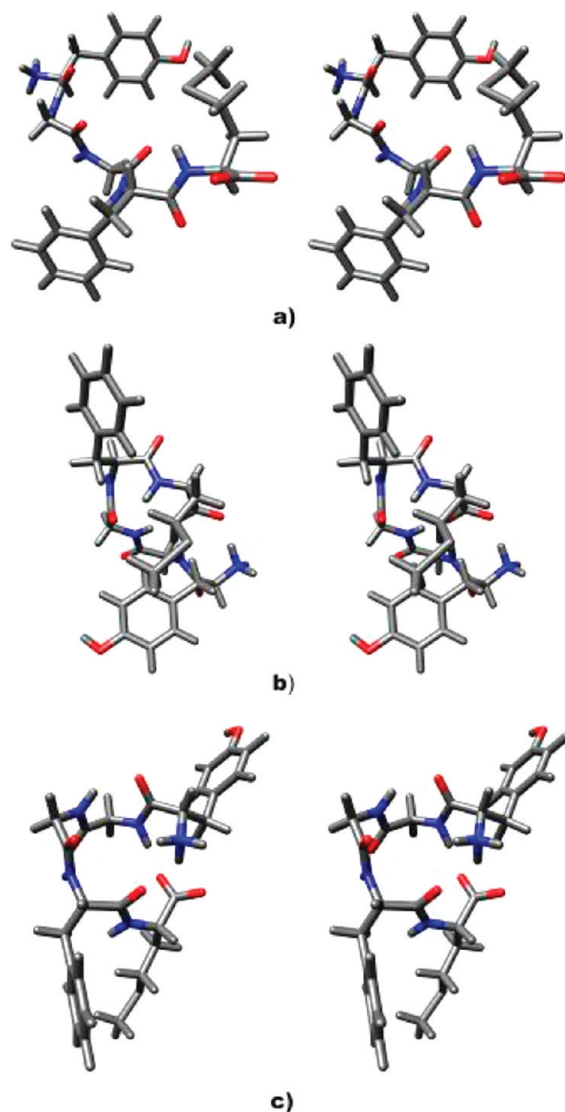


Figure 4. Stereo views of the lowest-energy MOLS (a) explicitly solvated, (b) implicitly solvated, and (c) vacuum structures. The energies of the explicitly solvated, implicitly solvated, and vacuum conformations are -350.6 , -80.9 , and -50.6 kcal/mol, respectively.

Table 1. Backbone Atoms RMSD (in Å) between MOLS Generated Lowest-Energy Structure and Other Reported Structures^a

	1PLW	1PLX	FABJIB	GEM
MOLS explicit solvent	2.08	2.13	1.60	2.75
MOLS implicit solvent	2.11	2.22	3.81	1.68
MOLS vacuum	2.59	2.52	3.83	1.37

^a NMR structures PDB³⁴ ID: 1PLW; 1PLX and X-ray crystal structure: CSD⁵ ID FABJIB; Global minimum energy structure:¹¹ GEM.

solvent, with the GEM and the experimental structures. It is evident that the solvated structures are, in general, closer to the experimental structures. Likewise, the nonsolvated structures are closer to the GEM. Once again, this matches expectations.

B. Structural Motifs. In order to study the effect of the solvent on the propensity of the molecule to form secondary structural motifs, we used the program PROMOTIF³⁵ to

Table 2. Best Sampled Structures^a and Corresponding Energies^b

	explicit solvent		implicit Solvent		vacuum	
	rmsd (Å)	energy (kcal/mol)	rmsd (Å)	energy (kcal/mol)	rmsd (Å)	energy (kcal/mol)
1PLW	0.73	-246.329	0.90	-70.502	1.01	65.845
1PLX	0.75	-321.723	0.82	-70.502	1.01	65.816
FABJIB	0.36	-308.425	0.33	-70.486	0.39	55.459
GEM	0.53	-218.652	0.69	-74.656	0.57	-39.414

^a Structures with the lowest rmsd as compared to either the GEM or the experimental structures. Rmsd is calculated for backbone atoms only. ^b NMR structures PDB³⁴ ID: 1PLW and 1PLX; X-ray crystal structure: CSD⁵ ID FABJIB; Global minimum energy structure:¹¹ GEM.

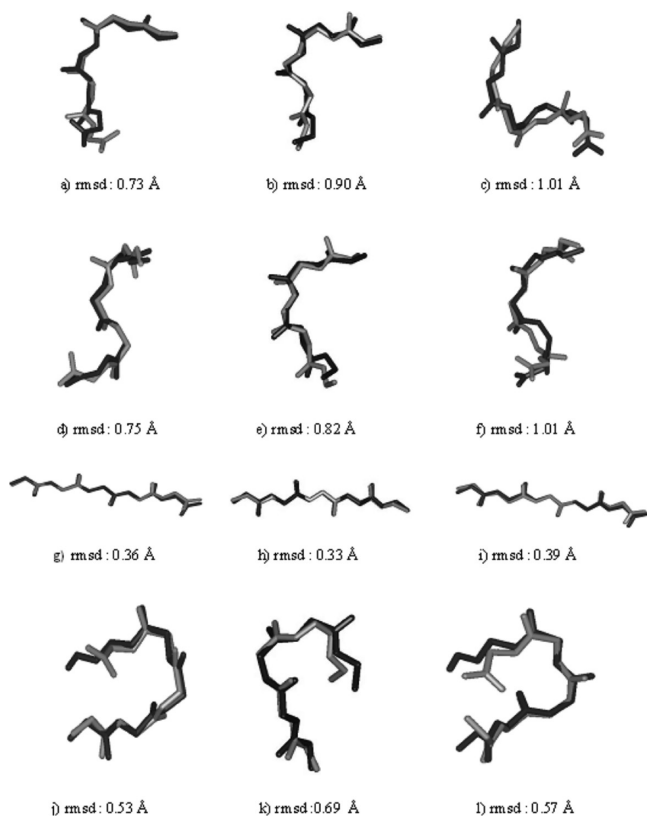


Figure 5. The superposition of the best predicted explicitly solvated, implicitly solvated, and vacuum structures with experimental and GEM structures. Figures (a–c) show superposition of the three structures, respectively, with the NMR structure 1PLW, (d–f) with the NMR structure 1PLX, (g–i) with the crystal structure CSD ID FABJIB, and (j–l) with the GEM structure. Native structures are dark (black), and MOLS structures are light (gray).

identify beta and gamma turns. Folded conformations with hydrogen bonds are termed as turns and the conformations without hydrogen bonds are termed bends.³⁶ (Both bends and turns are referred to by the general name ‘fold’.) A beta turn is defined for four consecutive residues (denoted by i , $i + 1$, $i + 2$, and $i + 3$), if the distance between the C^α atom of residue i and the C^α atom of residue $i + 3$ is less than 7 Å, and if the central two residues are not helical. A gamma turn is defined for three residues i , $i + 1$, and $i + 2$, if a hydrogen bond exists between residues i and $i + 2$, and the φ and ψ angles of residue $i + 1$ fall within 40° of one of

the two sets of values given in Table 4. The analysis shows that the 1 500 explicit solvated structures have a total of 543 beta motifs, of which 56 are beta turns, and the rest are beta bends. The 1 500 nonsolvated structures have 1 410 beta motifs with 82 beta turns and 1328 beta bends. It is apparent from these results that the presence of the water molecules reduces the number of intramolecular interactions, leading to a decrease in the number of secondary structural motifs.

Table 5 shows the distribution of the bends and turns among the various types of beta motifs. There are 32% of the solvated structures that have at least one beta bend or turn, and 75% of the nonsolvated structures have such a feature. Type IV folds constitute the large majority of the folds in both solvated and nonsolvated structures. It has been reported³⁷ that when the residue at the “ $i + 1$ ” and “ $i + 2$ ” positions is glycine, as in the present case, the peptide has a large propensity to adopt the type IV fold.

In both sets of generated structures, there are many single and double beta folds. There are also several structures with a beta and gamma fold, a beta fold and two gamma folds, two beta folds and a gamma fold, or two beta and gamma folds (Table 5). One of the well-folded solvated structures consists of one 1–4 type IV beta bend, one 2–5 type II’ beta bend, and a single 1–3 inverse gamma turn. This structure, shown in Figure 6, has a low energy of -204.7 kcal/mol. This structure also has one intramolecular hydrogen bond between Tyr¹ (CO) → Gly³(NH) and six intermolecular hydrogen bonds between the peptide and water molecules. However, it has a backbone rmsd of 1.37 Å with the GEM structure (and 3.45 Å with the crystal structure), indicating that the GEM structure may be different in the presence of solvent. The lowest-energy nonsolvated structure (-50.6 kcal/mol) has two type IV beta bends centered at Gly²–Gly³ and Gly³–Phe⁴ and double 1–3 inverse and 2–4 classic gamma turns. This structure, shown in Figure 7, has six hydrogen bonds with a backbone atom rmsd of 1.37 Å with the GEM structure and 3.83 Å with the crystal structure.

C. Hydrogen-Bonding Patterns. Intramolecular Hydrogen Bonds. The bonding patterns described by Mitsutake et al.³⁸ classify 12 possible hydrogen bonds between the backbone atoms in enkephalin. They are N1–O3, O1–N3, N1–O4, O1–N4, N1–O5, O1–N5, N2–O4, O2–N4, N2–O5, O2–N5, N3–O5, and O3–N5. Mitsutake classified Met-enkephalin structures into six classes of similar structures, referred to as C13, C14, C15, C24, C25, and C35. The class is denoted as C_{mn} where m and n are residues between which the hydrogen bond is formed. For example, a structure that belongs to class C14 would have at least one of the hydrogen bonds N1–O4 or N4–O1. Sometimes both are present giving rise to an antiparallel bonding pattern.

The distribution of intramolecular hydrogen bonds between backbone atoms is tabulated in Table 6. As may be expected, the numbers of such bonds in solvated structures are less than those observed in the nonsolvated structures. Among the solvated structures, the largest numbers of hydrogen bonds occur between residues 3 and 5 and among the nonsolvated structures between residues 1 and 5.

While most of the structures have only a single hydrogen bond, a few have more than one. There are three nonsolvated

Table 3. Total Number of MOLS Structures with Backbone Atoms RMSD Less than 1 Å with Experimental and GEM Structures

	explicit Solvent			implicit Solvent			vacuum		
	no. of structures	avg rmsd (Å)	avg energy (kcal/mol)	no. of structures	avg rmsd (Å)	avg energy (kcal/mol)	no. of structures	avg rmsd (Å)	avg energy (kcal/mol)
1PLW	4	0.88	-285.396	6	0.94	-67.740	nil	-	-
1PLX	3	0.86	-249.127	10	0.89	-69.874	nil	-	-
FABJIB	36	0.86	-277.480	97	0.85	-69.678	37	0.96	59.196
GEM	38	0.89	-211.532	1	0.69	-74.656	109	0.98	-38.103

Table 4. Gamma Turns

turn type	$\varphi (i + 1)$ (deg)	$\psi (i + 1)$ (deg)
classic	75.0	-64.0
inverse	-79.0	69.0

structures that have three hydrogen bonds each (Figure 8). In each case, the hydrogen bonds are between the oxygen atom of the terminal carbonyl group and the nitrogen atoms of residues 1, 2, and 3. These three structures have backbone atom rmsd's of 1.34, 1.23, and 1.10 Å, respectively, with

the GEM structure and 3.73, 3.66, and 3.58 Å with the crystal structure. These structures have hydrogen bonds between backbone atoms of adjacent residues (Phe⁴ (NH) → Met⁵ (OT1)).²⁸

Apart from the above, there are 10 solvated structures with main chain–side chain hydrogen bonds and 18 nonsolvated structures. Once again, the number is larger for the nonsolvated structures. These results show that, as may be expected, the solvated structures are closer to the experimental structures, which are extended, than to the theoretical GEM

Table 5. Distribution of Turns and Bends

a) Single Beta Turns and Bends															
type of beta motif	explicit					implicit					vacuum				
	turns		bends		total	turns		bends		total	turns		bends		total
	2-3 ^a	3-4	2-3	3-4		2-3	3-4	2-3	3-4		2-3	3-4	2-3	3-4	
I	1	1	-	1	3	-	-	-	-	-	-	-	-	2	2
II	2	-	2	-	4	1	-	-	-	1	-	-	1	-	1
VIII	1	-	1	7	9	-	-	1	2	3	-	-	2	9	11
I'	-	1	-	-	1	-	-	-	-	-	-	-	-	2	2
II'	2	4	3	1	10	1	-	-	-	1	3	4	3	4	14
IV	15	29	135	337	516	8	5	44	82	139	27	48	433	872	1 380
total	21	35	141	346	543	10	5	45	84	144	30	52	439	889	1 410

b) Single Gamma Turns												
type	explicit				implicit				vacuum			
	1-3 ^b	2-4	3-5	total	1-3	2-4	3-5	total	1-3	2-4	3-5	total
inverse	74	57	51	182	54	134	86	274	56	35	29	120
classic	63	97	24	184	72	99	25	196	57	186	61	304
total	137	154	75	366	126	233	111	470	113	221	90	424

c) Single and Double Folded Structures										
type	explicit			implicit				vacuum		
	single fold	double fold	total	single fold	double fold	triple fold	total	single fold	double fold	total
beta	427	58	485	136	4	-	140	834	288	1 122
gamma	298	34	332	381	40	3	424	340	42	382
total	725	92	817	517	44	3	564	1 174	330	1 504

d) Both Beta Folds and Gamma Turns				
type	no. of structures			
	explicit	implicit	vacuum	
single β -fold classic γ turn	77	39	179	
single β -fold inverse γ turn	38	23	58	
single β -fold double γ turns	10	13	18	
single β -fold triple γ turns	-	2	-	
double β -folds classic γ turn	13	2	82	
double β -folds inverse γ turn	2	-	18	
double β -folds double γ turns	7	-	24	

^a Residue number of central two residues of turn or bend. ^b Residue number of end residues of turn.

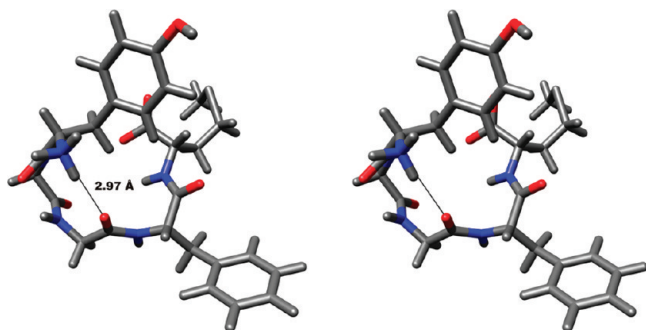


Figure 6. Stereo view of the explicitly solvated structure with two beta bends and an inverse gamma turn. This structure has a backbone rmsd of 1.37 Å with the GEM and 3.45 Å with the crystal structures.

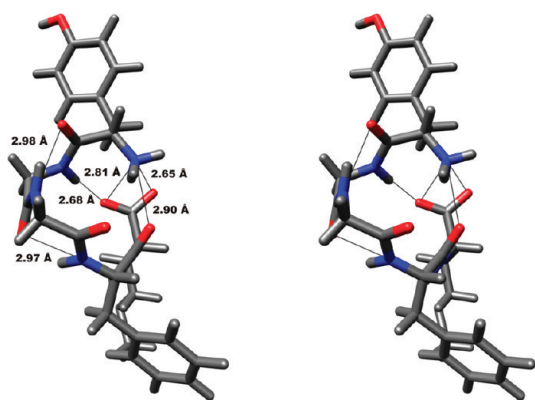


Figure 7. Stereo view of the lowest-energy vacuum structure with two beta bends and two gamma turns.

Table 6. Intramolecular Backbone Hydrogen Bond

type	explicit	implicit	vacuum
C13	2	1	27
C14	15	8	66
C15	59	11	275
C24	2	2	12
C25	61	11	181
C35	63	10	225
total	202	43	786

structure, which is folded. The situation is the reverse for the nonsolvated structures.

Peptide–water Hydrogen-Bonding Pattern. The number of water molecules bound to the peptide range from 5 to 15, with an average of eight waters in each, of the 1 500 solvated structures generated. In general, the extended structures, with lesser number of secondary structure motifs and intramolecular hydrogen bonds, have a greater number of hydrogen-bonded contacts with the water. Figure 9 shows an example. The folded structures, with a larger number of intramolecular hydrogen bonds, make fewer hydrogen-bonded contacts with water (Figure 10). Of the 1 500 generated structures, the one with the lowest energy has 12 hydrogen bonds between the peptide and the solvent water. The structure closest to the crystal structure has 10 peptide–water hydrogen bonds. It is clear that the intramolecular hydrogen bonds in the structures are destroyed by water, which in turn makes it closer to the experimental extended structures.

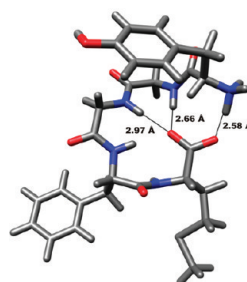
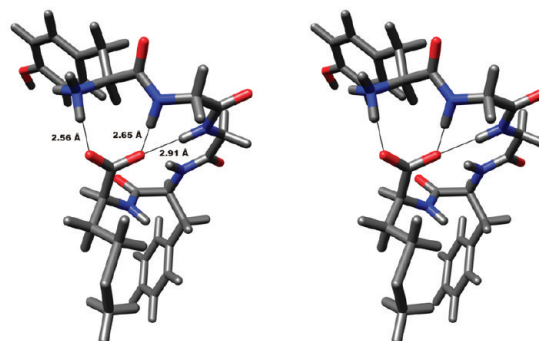
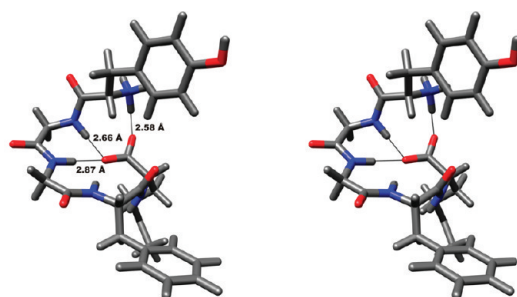


Figure 8. Stereo view of three vacuum structures with three backbone hydrogen bonds formed between residues 1–5, 2–5, and 3–5.

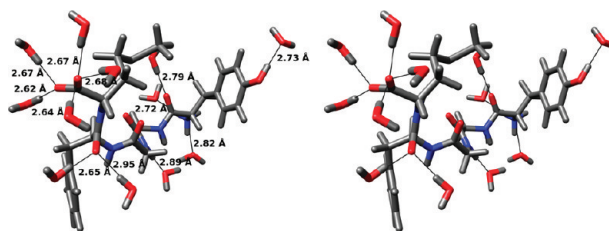


Figure 9. Stereo view of a structure without any intramolecular hydrogen bonds and with a large number of water molecules bonded to the peptide.

D. Energy Landscapes. The potential energy landscape of a peptide may be characterized by the set of locally stable conformations.³⁹ Since it is impossible to chart molecular energy landscapes in the full 3N-6 dimensional space, a smaller number of dimensions (usually two) is considered sufficient to capture the essential information required for energy landscape mapping, provided these dimensions are chosen carefully. The multidimensional conformational space can be reduced by principal component analysis (PCA)⁴⁰ or, more appropriately in the present context, the so-called principal coordinate analysis (PCoA).³⁹ In the present case, all the 1 500 structures generated in each set, solvated and

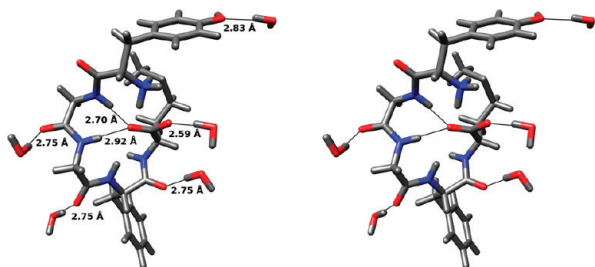


Figure 10. Stereo view of a structure with two intramolecular hydrogen bonds and few peptide–water hydrogen bonds.

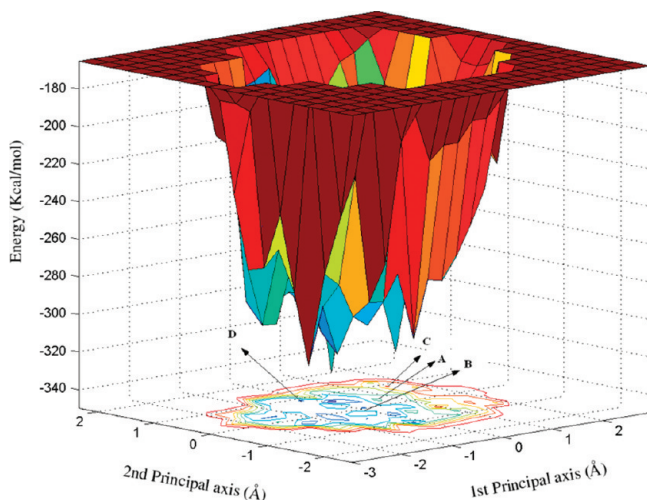


Figure 11. Energy landscape obtained for Met-enkephalin in the presence of explicit water. The two principal axes indicate conformational similarity, and the vertical axis indicates the relative energy.

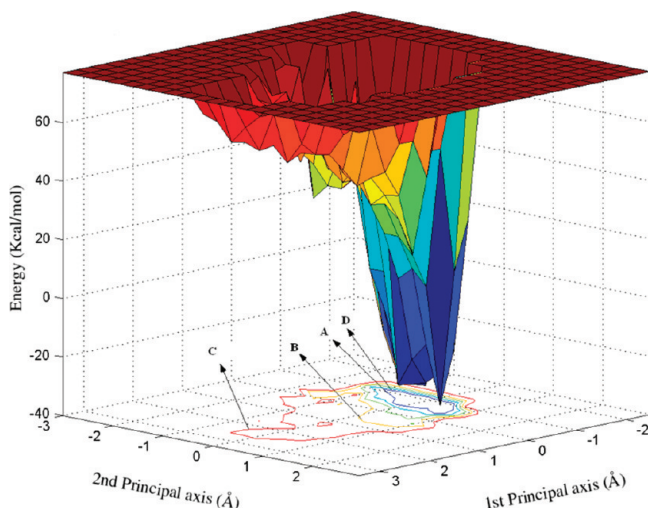


Figure 12. Energy landscape obtained for Met-enkephalin in the absence of water.

nonsolvated, were used to construct the respective energy landscapes. After the PCoorA, the first two principal coordinates accounted for 36.2% of the variance in the solvated set of structures and 49.1% of the variance in the nonsolvated set of structures. The energy landscapes for these three sets are shown in Figures 11 and 12. Clearly, solvation increases the width of the ‘folding funnel’⁴¹ as well as the ruggedness at the bottom of the funnel. There are many more

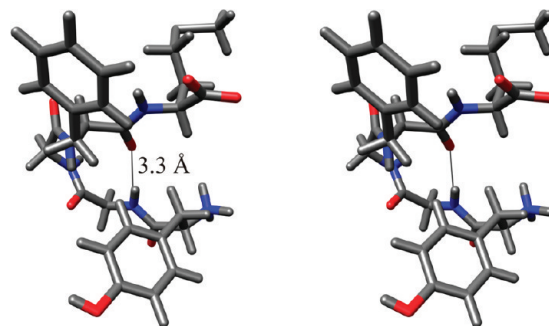


Figure 13. Stereo view of the implicitly solvated structure with two beta bends and a classic gamma turn. This structure has a backbone rmsd of 1.78 Å with the GEM and 3.46 Å with the crystal structures.

minima upon solvation than when the peptide is not solvated, though the energy differences between these minima are small. The respective energy minima are marked on the xy plane in both figures (the arrows labeled ‘A’), as are the points representing the experimental structures (arrows labeled ‘B’ for the NMR structure and ‘C’ for the X-ray crystal structure), and the GEM (arrow labeled ‘D’). In the energy landscape of the solvated structures the minimum-energy structure is closer to the experimental structures than to the GEM. It is vice versa in the landscape of nonsolvated structures.

E. Effects of the Implicit Solvent Model. The lowest-energy structure with implicit solvent has an energy of -80.9 kcal/mol. This structure is different from both the one generated with an explicit solvent (backbone rmsd of 2.94 Å) and the one generated in vacuum (1.99 Å). Table 1 shows that the lowest-energy implicit solvent structure is close to the GEM rather than the experimental structures, though among the 1 500 low-energy structures generated, there is a greater number closer to the experimental structures (Table 3).

Among these 1 500 structures, there are 144 beta motifs, of which 15 are beta turns and the rest are beta bends. There are 9% of the structures with an implicit solvent have at least one beta bend or turn. One of the well-folded structures consists of one 1–4 type IV beta bend, one 2–5 type IV beta bend, and one 2–4 classic gamma turn. This structure, shown in Figure 13, has an energy of -71.7 kcal/mol. It has a backbone rmsd of 1.78 Å with the GEM structure and 3.46 Å with the crystal structure.

The number of hydrogen bonds observed in these structures is even less than that observed in structures with an explicit solvent. The largest number of hydrogen bonds is between residues 1 and 5 or 2 and 5. There are three structures that have three hydrogen bonds each (Figure 14). In each case, the hydrogen bonds are between the oxygen atom of the terminal carbonyl group and the nitrogen atoms of residues 1, 2, and 3. These three structures have backbone rmsd of 1.09, 1.68, and 1.30 Å, respectively, with the GEM structure and 3.83, 3.81, and 3.77 Å with the crystal structure. Apart from the above, there are three structures with main chain–side chain hydrogen bonds.

All the 1 500 structures generated with an implicit solvent were used to construct the energy landscape. After the PCoorA procedure, the first two principal coordinates ac-

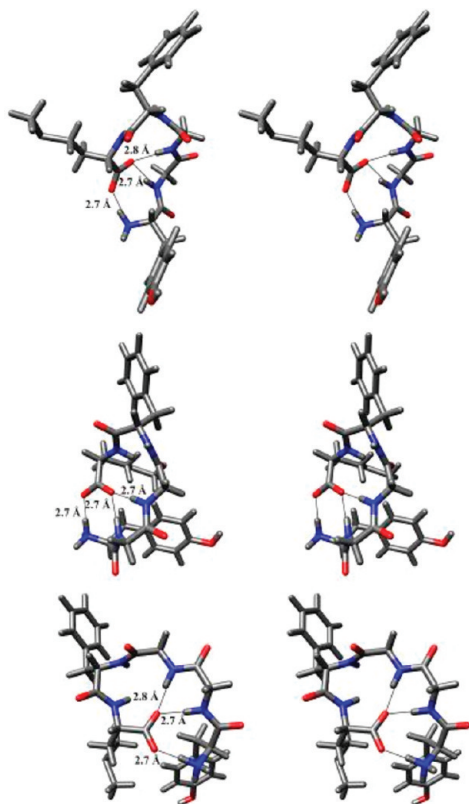


Figure 14. Stereo view of three implicitly solvated structures with three backbone hydrogen bonds formed between residues 1–5, 2–5, and 3–5.

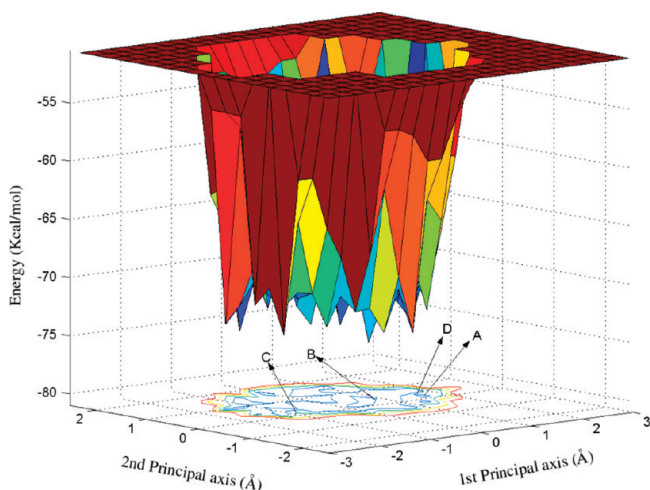


Figure 15. Energy landscape obtained for Met-enkephalin in the presence of implicit water.

counted for 27.4% of the conformational variance in the structures. Figure 15 shows the energy landscape drawn along these two principal coordinates. It may be seen that the effect of the implicit solvent are, in general, the same as that of an explicit solvent. Again, there are many more minima upon solvation than when the peptide is not solvated, though the energy differences between these minima are small.

Conclusion

We have used the MOLS algorithm to study the effect of solvation on the conformational energy landscape of Met-

enkephalin. The MOLS algorithm is particularly well suited for this study since it performs an exhaustive and unbiased sampling of the conformational space of the peptide to locate all the low-energy conformers. The study shows that inclusion of solvation is important to correctly model the conformational behavior of the molecule. Structures modeled in the presence of water molecules are far more similar to the experimental structures than when the water molecules are excluded.

The calculations also indicate that the pentapeptide Met-enkephalin prefers extended structures in an aqueous environment, as against tightly folded structures in the absence of water. Thus, the biologically relevant structure may be the extended structure as seen in the crystallographic and NMR experiments,^{5,34} rather than the GEM structures calculated by various workers.¹¹ The calculations are performed on the unbound molecule, whereas, as stated earlier in the Introduction, various workers⁴ have described a folded conformation for the molecule bound to the receptor protein. One may speculate that Met-enkephalin exists as an extended molecule in solution but assumes a folded form on binding to the receptor.

Acknowledgment. We thank the Department of Science and Technology, Government of India and UGC, Government of India under Centre of Advanced Study program for financial support.

References

- (1) Hughes, J.; Smith, T. W.; Kosterlitz, H. W.; Fothergill, L. A.; Morgan, B. A.; Morris, H. R. Identification of two related pentapeptides from the brain with potent opiate agonist activity. *Nature (London)*. **1975**, *258*, 577–579.
- (2) Schiller, P. W. Conformational analysis of enkephalin and conformation-activity relationships. In *The Peptides: Analysis, Synthesis, Biology*; Udenfriend, S., Meinhofer, J., Eds.; Academic Press: Orlando, FL, 1984; Vol. 6, pp 219–268.
- (3) Atweh, S. F.; Kuhar, M. J. Distribution and physiological significance of opioid receptors in the brain. *Br. Med. Bull.* **1983**, *39*, 47–52.
- (4) Ishida, T.; Kenmotsu, M.; Mino, Y.; Inoue, M.; Fujiwara, T.; Tomita, K.-I.; Kimura, T.; Sakakibara, S. X-ray diffraction studies of enkephalins. Crystal structure of [(4'-bromo) Phe 4, Leu 5] enkephalin. *Biochem. J.* **1984**, *218*, 677–689.
- (5) Griffin, J. F.; Langs, D. A.; Smith, G. D.; Blundell, T. L.; Tickle, I. J.; Bedarkar, S. The crystal structures of [Met⁵] enkephalin and a third form of [Leu⁵] enkephalin: Observations of a novel pleated β -sheet. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 3272–3276.
- (6) Graham, W. H.; Carter, E. S., II; Hicks, R. P. Conformational analysis of Met-enkephalin in both aqueous solution and in the presence of sodium dodecyl sulfate micelles using multidimensional NMR and molecular modeling. *Biopolymers*. **1992**, *32*, 1755–1764.
- (7) Montcalm, T.; Cui, W.; Zhao, H.; Guarnieri, F.; Wilson, S. R. Simulated annealing of Met-enkephalin: low energy states and their relevance to membrane-bound, solution and solid-state conformations. *J. Mol. Struct. (THEOCHEM)* **1994**, *308*, 37–51.
- (8) Eisenmenger, F.; Hansmann, U. H. E. Variation of the energy landscape of a small peptide under a change from the ECEPP/2

- force field to ECEPP/3. *J. Phys. Chem. B* **1997**, *10*, 3304–3310.
- (9) Hansmann, U. H. E.; Masuya, M.; Kamoto, Y. Characteristic temperatures of folding of a small peptide. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 10652–10656.
- (10) Zhan, L.; Chen, J. Z. Y.; Liu, W. Conformational study of Met-enkephalin based on the ECEPP force fields. *Biophys. J.* **2006**, *91*, 2399–2404.
- (11) Isogai, Y.; Nemethy, G.; Scheraga, H.A. Enkephalin: conformational analysis by means of empirical energy calculations. *Proc. Natl. Acad. Sci. U.S.A.* **1977**, *74*, 414–418.
- (12) Klepeis, J. L.; Floudas, C. A. Free-energy calculations for peptides via deterministic global optimization. *J. Chem. Phys.* **1999**, *110*, 7491–7512.
- (13) Carlucci, L. Conformational analysis of [Met⁵]-enkephalin: solvation and ionization considerations. *J. Comput.-Aided Mol. Design* **1998**, *12*, 195–213.
- (14) Ooi, T.; Oobatake, M.; Nemethy, G.; Scheraga, H. A. Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 3086–3090.
- (15) Shen, M.-Y.; Freed, K. F. Long time dynamics of Met-enkephalin: comparison of explicit and implicit solvent models. *Biophys. J.* **2002**, *82*, 1791–1808.
- (16) Zaman, M. H.; Shen, M.-Y.; Berry, R. S.; Freed, K. F. Computer simulation of Met-enkephalin using explicit atom and united atom potentials: similarities, differences, and suggestions for improvement. *J. Phys. Chem. B* **2003**, *107*, 1685–1691.
- (17) Vengadesan, K.; Gautham, N. Enhanced sampling of the molecular potential energy surface using mutually orthogonal Latin squares: Application to peptide structures. *Biophys. J.* **2003**, *84*, 2897–2906.
- (18) Vengadesan, K.; Gautham, N. Conformational studies on enkephalins using the MOLS technique. *Biopolymers* **2004**, *74*, 476–494.
- (19) Vengadesan, K.; Gautham, N. Energy landscape of Met-enkephalin and Leu-enkephalin drawn using mutually orthogonal Latin squares sampling. *J. Phys. Chem. B* **2004**, *108*, 11196–11205.
- (20) Olszewski, K. A.; Piela, L.; Scheraga, H. A. Mean field theory as a tool for intramolecular conformational optimization. I. Tests on terminally-blocked alanine and Met-enkephalin. *J. Phys. Chem.* **1992**, *96*, 4672–4676.
- (21) Koehl, P.; Delarue, M. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J. Mol. Biol.* **1994**, *239*, 249–275.
- (22) Finney, D. J. Randomized blocks and Latin squares. In *Experimental Design and its Statistical Basis*; Cambridge University Press: London, U.K., 1955; pp 45–67.
- (23) Ito, K. Latin squares. In *Encyclopedic Dictionary of Mathematics*; MIT Press: Cambridge, MA, 1987; Vol. 2, pp 891–892.
- (24) Cochran, W. G.; Cox, G. M. *Experimental Designs*, 2nd ed.; John Wiley and Sons: New York, 1957; pp 117–133.
- (25) Fisher, R. A. *The Design of Experiments*, 7th ed.; Oliver and Boyd: London, 1960; pp 70–92.
- (26) Ryser, H. J. *Combinatorial Mathematics*, Mathematical Association of America: Washington, DC, 1963; pp 79–84.
- (27) Liu, C. L. *Introduction to Combinatorial Mathematics*, McGraw-Hill Book Company: New York, 1968; pp 1–22.
- (28) Kriz, Z.; Carlsen, P. H. J.; Koca, J. Conformational features of linear and cyclic enkephalins. A computational study. *J. Mol. Struct. (THEOCHEM)* **2001**, *540*, 231–250.
- (29) Humphrey, W.; Dalke, A.; Schulten, K. VMD - Visual molecular dynamics. *J. Mol. Graph.* **1996**, *14*, 33–38.
- (30) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L., Jr.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E., III; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; 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*, 3586–3616.
- (31) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.
- (32) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781–1802.
- (33) Schaefer, M.; Karplus, M. A comprehensive analytical treatment of continuum electrostatics. *J. Phys. Chem.* **1996**, *100*, 1578–1599.
- (34) Marcotte, I.; Separovic, F.; Auger, M.; Gagne, S. M. A multidimensional 1H NMR investigation of the conformation of methionine-enkephalin in fast-tumbling. *Biophys. J.* **2004**, *86*, 1587–1600.
- (35) Hutchinson, E. G.; Thornton, J. M. PROMOTIF - A program to identify and analyze structural motifs in proteins. *Protein Sci.* **1996**, *5*, 212–220.
- (36) Van der Spoel, D.; Berendsen, H. J. C. Molecular dynamics simulations of Leu-enkephalin in water and DMSO. *Biophys. J.* **1997**, *72*, 2032–2041.
- (37) Lewis, P. N.; Momany, F. A.; Scheraga, H. A. Chain reversals in proteins. *Biochim. Biophys. Acta* **1973**, *303*, 211–229.
- (38) Mistsutake, A.; Hansmann, U. H. E.; Okamoto, Y. Temperature dependence of distributions of conformations of a small peptide. *J. Mol. Graphics Modell.* **1998**, *16*, 226–238.
- (39) Levy, Y.; Becker, O. M. Energy landscapes of conformationally constrained peptides. *J. Chem. Phys.* **2001**, *114*, 993–1009.
- (40) Becker, O. M.; Levy, Y.; Ravitz, O. Flexibility, conformation spaces and bioactivity. *J. Phys. Chem. B* **2000**, *104*, 2123–2135.
- (41) Garcia, A. E.; Sanbonmatsu, K. Y. Exploring the energy landscape of a β hairpin in explicit solvent. *Proteins: Struct., Funct., Genet.* **2001**, *42*, 345–354.

Erratum

Perspective on Diabatic Models of Chemical Reactivity as Illustrated by the Gas-Phase S_N2 Reaction of Acetate Ion with 1,2-Dichloroethane.

[*J. Chem. Theory Comput.* 5, 1–22 (2009)]. By Rosendo Valero, Lingchun Song, Jiali Gao,* and Donald G. Truhlar*.

Pages 1–22. Our article¹ compared five quantum mechanical methods for computing diabatic states for the gas-phase haloalkane dehalogenase (Dh1A) model reaction between an acetate ion and dichloroethane to one another and to the molecular mechanical diabatic states of EVB theory. The most recent application of EVB to Dh1A is the work of Rosta et al.² This paper says that the EVB parameters are a small modification of those used in Olsson et al.³ The Olsson et al. paper says that the EVB parameters they used are a modification of those reported by Shurki et al.⁴ Neither paper reports the modified parameters. Reference 4 states that the EVB parameters used for reaction in water and in the protein are given in the Supporting Information, and we therefore chose that paper as one where we would have a full set of parameters. The parameters we used—including corrections of typographical errors and a correction of a factor of 2 to the exponential parameters of the nonbonded van der Waals interactions that was kindly provided by the first author of ref 4—are included in our paper.¹ (Note the following typos in our Supporting Information: all θ_0 values except that for O–C–O were set equal to 109.5 deg, as in ref 4). We did not include any terms whose parameters are not given in ref 4.

The parameters in ref 4 include a single gas-phase shift parameter α without explicitly specifying to which diagonal element it should be added. In our calculation,¹ we added this to the first EVB diabatic state because this yielded an

energy difference (–7 kcal/mol) between the product and reactant states in reasonable accord with the most accurate values reported in ref 1 (–11 kcal/mol from G3SX and M06-2X calculations). Files recently posted on the Warshel group Web site⁵ (the files are dated Feb. 23, 2009) indicate that they shifted the second diabatic state. Shifting the second diabatic state in our calculations yielded a very endothermic reaction, which is unreasonable. We conclude that additional terms—not included in our calculations—whose parameters were not published are needed to reproduce the potential underlying the published EVB calculations of refs 2–4. We therefore retract all quantitative results presented in our paper for the EVB potential of ref 4.

The removal of the molecular mechanical EVB calculations has no effect on the larger portion of the paper that is devoted to the quantum mechanical diabatic surfaces obtained by the 4-fold way and MOVb.

References

- (1) Valero, R.; Song, L.; Gao, J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2009**, *5*, 1–22.
- (2) Rosta, E.; Klähn, M.; Warshel, A. *J. Phys. Chem. B* **2006**, *110*, 2934–2941.
- (3) Olsson, M. H. M.; Warshel, A. *J. Am. Chem. Soc.* **2004**, *126*, 15167–15179.
- (4) Shurki, A.; Strajbl, M.; Villa, A.; Warshel, A. *J. Am. Chem. Soc.* **2002**, *124*, 4097–4107.
- (5) <http://futura.usc.edu/wgroup/evb.html> (accessed Mar. 9, 2009).
CT9002459

10.1021/ct9002459

Published on Web 06/10/2009

JCTC

Journal of Chemical Theory and Computation

Extracting Kinetic and Stationary Distribution Information from Short MD Trajectories via a Collection of Surrogate Diffusion Models. [*J. Chem. Theory Comput.* 5, 47–58 (2009)]. By Christopher P. Calderon* and Karunesh Arora.

Pages 47–58. There were minor errors, e.g. colons appear where decimal points should be and “MO” replaced “MD”, in a few portions of the Results.

However, one scientifically relevant typographical error in eq 6 of the print version is worth noting. The integral term should have been in the argument of the exponential. The corrected version of this equation should read

$$p^{EQ}(\Phi; \mathcal{G}) = \frac{Z}{(\sigma^{\mathcal{G}}(\Phi))^2} \exp\left(\int_{\Phi_{\text{REF}}}^{\Phi} \mu^{\mathcal{G}}(\Phi') / (\sigma^{\mathcal{G}}(\Phi'))^2 d\Phi'\right) \quad (6)$$

The results reported used the correct expression.

CT900321J

10.1021/ct900321j

Published on Web 07/07/2009